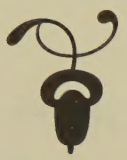
William P. Alston ▮ Michael Bergmann
Evan Fales ▮ Jerry Fodor
Trenton Merricks ▮ Timothy O'Connor
Richard Otte ▮ Alvin Plantinga
William Ramsey ▮ Ernest Sosa
W. J. Talbott ▮ James Van Cleve
Edited by JAMES BEILBY
(with a response by Alvin Plantinga)

NATURALISM DEFEATED?

ESSAYS ON PLANTINGA'S EVOLUTIONARY ARGUMENT AGAINST NATURALISM

Naturalism Defeated?



Naturalism Defeated?

Essays on Plantinga's Evolutionary Argument against Naturalism

Edited by
James Beilby

Cornell University Press
Ithaca & London

Copyright © 2002 by Cornell University

All rights reserved. Except for brief quotations in a review, this book, or parts thereof, must not be reproduced in any form without permission in writing from the publisher. For information, address Cornell University Press, Sage House, 512 East State Street, Ithaca, New York 14850.

First published 2002 by Cornell University Press
First printing, Cornell Paperbacks, 2002

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Naturalism defeated? : essays on Plantinga's evolutionary argument against naturalism / edited by James Beilby.
p. cm.
Includes bibliographical references and index.
ISBN 0-8014-3972-8 (cloth : alk. paper)—ISBN 0-8014-8763-3 (pbk. : alk. paper)
1. Plantinga, Alvin. 2. Naturalism. 3. Human evolution. 4. Knowledge, Theory of. I. Beilby, James K.
B931.P454 N38 2002
210'.92—dc21

2001006111

Cornell University Press strives to use environmentally responsible suppliers and materials to the fullest extent possible in the publishing of its books. Such materials include vegetable-based, low-VOC inks and acid-free papers that are recycled, totally chlorine-free, or partly composed of nonwood fibers. For further information, visit our website at www.cornellpress.cornell.edu.

Chapter 2, "Is Science Biologically Possible?" originally appeared in Jerry Fodor, *In Critical Condition* (Cambridge: MIT Press, 1998), 189–202. Reprinted with permission of the author and publisher.

Cloth printing 10 9 8 7 6 5 4 3 2 1
Paperback printing 10 9 8 7 6 5 4 3 2 1

Contents

Preface vii
Introduction: The Evolutionary Argument against Naturalism *1*
Alvin Plantinga

Part I Science and Evolution

Naturalism Defended *15*
William Ramsey

Is Science Biologically Possible? *30*
Jerry Fodor

Darwin's Doubt, Calvin's Calvary *43*
Evan Fales

Part II Skepticism

Commonsense Naturalism *61*
Michael Bergmann

Plantinga's Evolutionary Meditations *91*
Ernest Sosa

Can Atheists Know Anything? *103*
James Van Cleve

Part III Conditional Probabilities and Confirmation Theory

A House Divided against Itself Cannot Stand *129*
Timothy O'Connor

Conditional Probabilities in Plantinga's Argument *135*
Richard Otte

Part IV The Nature of Epistemic Defeat

The Illusion of Defeat *153*
W. J. Talbott

Conditional Probability and Defeat *165*
Trenton Merricks

Plantinga, Naturalism, and Defeat *176*
William Alston

Reply to Beilby's Cohorts *204*
Alvin Plantinga

Contributors *277*
Index *281*

Preface

Throughout his academic career Alvin Plantinga has defended the intellectual credibility of Christian theism. Recently, he has expanded his defense to include a bold and controversial claim: Not only is theistic belief rational, but one who denies the existence of a creative deity and accepts contemporary evolutionary theory is *irrational* in doing so. More accurately, the conjunction of metaphysical naturalism (N)—namely, the view that only natural objects, kinds, and properties are real—and evolution (E) is, according to Plantinga, self-defeating. Those who accept both N and E have a "defeater" for the belief that human cognitive faculties, so evolved, are reliable. This defeater, according to Plantinga, cannot itself be defeated and thereby constitutes a defeater for any belief produced by those cognitive faculties, including the beliefs which comprise N&E. Therefore, despite the fact that metaphysical naturalism and evolution are typically thought of as very closely and comfortably connected, taken together, their conjunction cannot rationally be held. Hume's game of backgammon beckons.

Plantinga's argument should *not* be mistaken for an argument against evolutionary theory in general or, more specifically, against the claim that humans might have evolved from more primitive life forms. Rather, the purpose of his argument is to show that the denial of the existence of a creative deity is problematic. It is *the conjunction* of naturalism and evolution that suffers from the crippling deficiency of self-defeat, a deficiency not shared by the conjunction of *theism* and current evolutionary doctrine.

One of the most fascinating aspects of Plantinga's evolutionary argument against naturalism is its multilayered nature. It raises issues of interest to epistemologists, philosophers of mind, evolutionary biologists, and philosophers of religion, just to name a few. The diversity of philosophical questions raised

by Plantinga's argument also contributes to its complexity. It is the purpose of this volume not only to consider this fascinating argument but to shed some light on the multifaceted and important interaction between these different philosophical subdisciplines. To better understand both Plantinga's argument and the organization of the essays in this volume, let's take a closer look at the structure of Plantinga's argument itself.

Plantinga's evolutionary argument involves three steps. First, Plantinga claims that the objective conditional probability that we have reliable cognitive faculties, given naturalism and evolutionary theory, is either low or, since it is difficult to even start to specify the relevant probabilities, inscrutable. Hence:

(1) P(R/N&E) is either low or inscrutable [where R stands for the proposition: "Human cognitive faculties are reliable"]

According to Plantinga, the mechanisms of evolution select for adaptive behavior, not necessarily true belief, and it is not obvious that adaptive behavior guarantees, or even makes probable, true belief. Evaluating the first step of Plantinga's argument involves considering the nature of evolutionary mechanisms and the nature of the relationship between belief and behavior from an evolutionary point of view. Part 1 addresses these kinds of questions.

The second step of Plantinga's argument involves the claim that one who accepts N&E and comes to realize the truth of (1) acquires a defeater for R. Hence:

(2) If S accepts N&E and (1), she has a rationality defeater for her belief in R

Even the inscrutability of P(R/N&E), according to Plantinga, is sufficient to give one who accepts N&E a reason to withhold belief in R. Further, this defeater cannot itself be defeated since any prospective defeater-defeater would involve beliefs which would be subject to defeat as well.

It seems clear that if the naturalist, or anyone for that matter, came to believe that she had a defeater for R, then the third and final step of Plantinga's argument would certainly follow:

(3) S has a defeater for all of her beliefs, one of which is N&E

But is (2) true? Assessing the truth value of (2) is a complex matter, one which raises many important philosophical questions. First of all, there is the question of skepticism, what it is and how it is to be avoided. These questions, and the contributions of René Descartes and Thomas Reid to the subject, will be discussed in part 2. Second, evaluating (2) will involve addressing what Plantinga calls "the Conditionalization Problem." Probability judgments depend crucially on the information conditionalized on. Plantinga's argument

makes use of the probability of R conditional on N&E. Is this justified? These matters are addressed in part 3. Finally, there is the complex but extremely important matter of the nature of "defeaters." The essays in part 4 shed some light on this topic as it relates to Plantinga's argument.

Of course, many of the essays do not fit neatly into a single category, and further, there undoubtedly is overlap between the categories themselves. (Don't all arguments about skepticism ultimately involve defeaters?) This is testimony to the complex web of philosophical questions invoked by Plantinga's argument. The grouping of the essays into quasi-topical categories is merely an attempt to enhance understanding by providing some structure.

Plantinga's reply to the essays begins with an in-depth discussion of several topics crucial to his argument. He then responds to each of the essays in such a way as to be able to refer to earlier essays as he proceeds.

Finally, some attention should be given to the antecedents of the material in this volume. First of all, Plantinga's evolutionary argument bears affinities to the arguments of C. S. Lewis in *Miracles* (especially chapters 3 and 13) and Richard Taylor in *Metaphysics* (chapter 10), although neither develops their arguments nearly as fully as Plantinga has. For Plantinga's part, this argument has appeared in "An Evolutionary Argument against Naturalism," *Logos* 12 (1991), and in chapter 12 of *Warrant and Proper Function* (Oxford University Press, 1993). In the years since, a number of articles have addressed Plantinga's argument, and Plantinga has responded to some of them in *Warranted Christian Belief* (Oxford University Press, 2000) and "Naturalism Defeated," an unpublished paper.[1] "Naturalism Defeated," despite being unpublished, has been widely circulated, and many of the essays in this volume make explicit use of it when discussing Plantinga's argument. Since "Naturalism Defeated" is not in print, however, the contributors to this volume have sought to make their references to it "stand alone."

No project worth its salt comes to fruition in a vacuum. I have received invaluable advice and support from many people during the course of my work on this volume. Thanks go to Mike Kukuska, who kindly prepared the index. I would be remiss if I failed to mention at least a few of those to whom I am indebted. Tony Peressini has been a valuable conversation partner on this topic, furthering my understanding of many aspects of Plantinga's argument. In addition, in the early stages of this project, Tony was a coeditor, and his suggestions helped shape the structure and makeup of this volume. I also owe an immeasurable debt of gratitude to my family. My children, Sierra, Madeline, and Zachary, are an unending source of joy, inspiration, and humor, and I am particularly grateful to my wife, Michelle. She has sustained and encouraged me in all of my endeavors, academic and otherwise. Finally, I wish to ex-

[1] Plantinga has also responded to an essay on this topic by Keith Lehrer in "Respondeo," in *Warrant in Contemporary Epistemology*, ed. Jonathan Kvanvig (Lanham, Md.: Rowman and Littlefield, 1996), 333–338.

press my thanks to Al Plantinga. In college I read his *God and Other Minds* (Cornell University Press, 1967). His clarity of thought and expression coupled with his characteristic wit were largely responsible for my burgeoning interest in philosophy and theology. Since then I have greatly enjoyed and appreciated his gracious advice, tutelage, and friendship (not to speak of the disc golf matches). I gratefully dedicate this book to him.

Jim Beilby

St. Paul, Minnesota

Naturalism Defeated?

Introduction

The Evolutionary Argument against Naturalism: An Initial Statement of the Argument

Alvin Plantinga

It's a delight to express my gratitude to James Beilby and his cohorts for these probing and challenging essays on my evolutionary argument against naturalism (hereafter EAAN). I have learned much from them, not only about EAAN, but also about defeaters and about warrant itself.

I proposed EAAN in the last chapter of *Warrant and Proper Function*.[1] Since I suppose it's at least logically possible that you do not have a copy of *Warrant and Proper Function* on your desk at the moment, I'll begin by briefly restating the argument.[2] Take *philosophical naturalism* to be the belief that there aren't any supernatural beings—no such person as God, for example, but also no other supernatural entities, and nothing at all like God.[3] My claim was that naturalism and contemporary evolutionary theory are at serious odds with one another—and this despite the fact that the latter is ordinarily thought to be one of the main pillars supporting the edifice of the former. (Of course I am *not* attacking the theory of evolution, or the claim that human beings have evolved from simian ancestors, or anything in that neighborhood; I am instead attacking the conjunction of *naturalism* with the view that human beings have evolved in that way. I see no similar problems with the conjunction of *theism* and the idea that human beings have evolved in the way contemporary

[1] New York: Oxford University Press, 1993.

[2] But first I must acknowledge an error: at a certain point the statement of the "preliminary argument" (*Warrant and Proper Function*, 228–229; hereafter *WPF*) confuses the unconditional objective or logical probability of R with its probability conditional on our background knowledge. For details and repair, see *Warranted Christian Belief* (New York: Oxford University Press, 2000), 229ff. The main argument, happily, is unaffected.

[3] If my project were giving an analysis of philosophical naturalism, much more would have to be said (for example, if we don't know what naturalism is, will it help to explain it in terms of supernaturalism?); for present purposes we can ignore the niceties.

evolutionary science suggests.) More particularly, I argued that the conjunction of naturalism with the belief that we human beings have evolved in conformity with current evolutionary doctrine—'evolution' for short—is in a certain interesting way self-defeating or self-referentially incoherent. Still more particularly, I argued that naturalism and evolution—'N&E' for short—furnishes one who accepts it with a *defeater* for the belief that our cognitive faculties are reliable—a defeater that can't be defeated. But then this conjunction also furnishes a defeater for any belief produced by our cognitive faculties, including, in the case of one who accepts it, N&E itself: hence its self-defeating character.

The Argument

So much for a quick overview of the argument. More specifically, EAAN begins from certain doubts about the *reliability* of our cognitive faculties, where, roughly, a cognitive faculty—memory, perception, reason—is reliable if the great bulk of its deliverances are true.[4] These doubts are connected with the *origin* of our cognitive faculties. According to current evolutionary theory, we human beings, like other forms of life, have developed from aboriginal unicellular life by way of such mechanisms as natural selection and genetic drift working on sources of genetic variation: the most popular is random genetic mutation. Natural selection discards most of these mutations (they prove deleterious to the organisms in which they appear), but some of the remainder turn out to have adaptive value and to enhance fitness; they spread through the population and thus persist. According to this story, it is by way of these mechanisms, or mechanisms very much like them, that all the vast variety of contemporary organic life has developed; and it is by way of these same mechanisms that our cognitive faculties have arisen.

Now according to traditional Christian (and Jewish and Muslim) thought, we human beings have been created in the image of God. This means, among other things, that God created us with the capacity for achieving *knowledge*—knowledge of our environment by way of perception, of other people by way of something like what Thomas Reid calls *sympathy*, of the past by memory and testimony, of mathematics and logic by reason, of morality, of our own mental life, of God himself, and much more.[5] And the above evolutionary ac-

[4] *Very* roughly: a thermometer stuck on seventy-two degrees isn't reliable even if it is located somewhere (San Diego?) where it is seventy-two degrees nearly all of the time. What the thermometer (and our cognitive faculties) would do if things were different in certain (hard to specify) respects is also relevant. Again, if our aim were to analyze *reliability* much more would have to be said. Note that for reliability thus construed, it is not enough that the beliefs produced be fitness enhancing.

[5] Thus Thomas Aquinas: "Since human beings are said to be in the image of God in virtue of their having a nature that includes an intellect, such a nature is most in the image of God in virtue of being most able to imitate God" (*Summa Theologica* Ia q. 93 a. 4); and "Only in rational creatures is there found a likeness of God which counts as an image. . . . As far as a likeness of the

count of our origins is compatible with the theistic view that God has created us in his image.[6] So evolutionary theory taken by itself (without the patina of philosophical naturalism that often accompanies expositions of it) is not as such in tension with the idea that God has created us and our cognitive faculties in such a way that the latter are reliable, that (as the medievals liked to say) there is an adequation of intellect to reality.

But if *naturalism* is true, there is no God, and hence no God (or anyone else) overseeing our development and orchestrating the course of our evolution. And this leads directly to the question whether it is at all likely that our cognitive faculties, given naturalism and given their evolutionary origin, would have developed in such a way as to be reliable, to furnish us with mostly true beliefs. Darwin himself expressed this doubt: "With me," he said,

> the horrid doubt always arises whether the convictions of man's mind, which has been developed from the mind of the lower animals, are of any value or at all trustworthy. Would any one trust in the convictions of a monkey's mind, if there are any convictions in such a mind?[7]

The same thought is put more explicitly by Patricia Churchland. She insists that the most important thing about the human brain is that it has evolved; this means, she says, that its principal function is to enable the organism to *move* appropriately:

> Boiled down to essentials, a nervous system enables the organism to succeed in the four F's: feeding, fleeing, fighting and reproducing. The principle chore of nervous systems is to get the body parts where they should be in order that the organism may survive. . . . Improvements in sensorimotor control confer an evolutionary advantage: a fancier style of representing is advantageous *so long as it is*

---

divine nature is concerned, rational creatures seem somehow to attain a representation of [that] type in virtue of imitating God not only in this, that he is and lives, but especially in this, that he understands" (*Summa Theologica* Ia q.93 a.6).

[6] You might think not: if our origin involves *random* genetic variation, then we and our cognitive faculties would have developed by way of *chance* rather than by way of design, as would be required by our having been created by God in his image. But this is to import far too much into the biologist's term 'random'. Those random variations are random in the sense that they don't arise out of the organism's design plan and don't ordinarily play a role in its viability; perhaps they are also random in the sense that they are not predictable. But of course it doesn't follow that they are random in the much stronger sense of not being caused, orchestrated, and arranged by God. And suppose the biologists, or others, *did* intend this stronger sense of 'random': then their theory (call it 'T') would indeed entail that human beings have not been designed by God. But T would not be more probable than not with respect to the evidence. For there would be an empirically equivalent theory (the theory that results from T by taking the weaker sense of 'random' and adding that God has orchestrated the mutations) that is inconsistent with T but as well supported by the evidence; if so, T is not more probable than not with respect to the relevant evidence.

[7] Letter to William Graham, Down, July 3, 1881, in *The Life and Letters of Charles Darwin Including an Autobiographical Chapter*, ed. Francis Darwin (London: John Murray, Albermarle Street, 1887), 1:315–316. Evan Fales and Omar Mirza have pointed out that Darwin probably had in mind, here, not everyday beliefs such as that the teapot is in the cupboard, but something more like religious and philosophical convictions.

> *geared to the organism's way of life and enhances the organism's chances of survival* [Churchland's emphasis]. Truth, whatever that is, definitely takes the hindmost.[8]

What Churchland means, I think, is that evolution is directly interested (so to speak) only in *adaptive behavior* (in a broad sense including physical functioning), not in true belief. Natural selection doesn't care what you *believe*; it is interested only in how you *behave*. It selects for certain kinds of behavior: those that enhance fitness, which is a measure of the chances that one's genes will be widely represented in the next and subsequent generations. It doesn't select for belief, except insofar as the latter is appropriately related to behavior. But then the fact that we have evolved guarantees at most that we *behave* in certain ways—ways that contribute to our (or our ancestors') surviving and reproducing in the environment in which we have developed. Perhaps Churchland's claim can be understood as the suggestion that the objective probability[9] that our cognitive faculties are reliable, given naturalism and given that we have been cobbled together by the processes to which contemporary evolutionary theory calls our attention, is low. Of course she doesn't explicitly mention naturalism, but it certainly seems that she is taking it for granted. For if theism were true, God might be directing and orchestrating the variation in such a way as to produce, in the long run, beings created in his image and thus capable of knowledge; but then it wouldn't be the case that truth takes the hindmost.

We can put Churchland's claim as

P(R/N&E) is low,

where 'R' is the proposition that our cognitive faculties are reliable, 'N' the proposition that naturalism is true, and 'E' the proposition that we have evolved according to the suggestions of contemporary evolutionary theory.[10] I believe this thought—the thought that P(R/N&E) is low—is also what worries Darwin in the above quotation: I shall therefore call it 'Darwin's Doubt'.

Are Darwin and Churchland right? Well, they are certainly right in thinking that natural selection is directly interested only in behavior, not belief, and that it is interested in belief, if at all, only indirectly, by virtue of the relation between behavior and belief. If adaptive behavior guarantees or makes probable reliable faculties, then perhaps P(R/N&E) will be fairly high: we or rather our ancestors engaged in at least reasonably adaptive behavior, so it must be that our cognitive faculties are at least reasonably reliable, in which

[8] Churchland, "Epistemology in the Age of Neuroscience," *Journal of Philosophy* 84 (October 1987): 548.

[9] For an account of objective probability, see *WPF*, 161ff.

[10] In *WPF* the probability at issue was the slightly more complex P(R/N&E&C), where C was a proposition setting out some of the main features of our cognitive system (see *WPF*, 220). I now think the additional complexity unnecessary.

case it is likely that most of our beliefs are true. On the other hand, if our having reliable faculties *isn't* guaranteed by or even particularly probable with respect to adaptive behavior, then presumably P(R/N&E) will be rather low. If, for example, behavior isn't caused or governed by belief, the latter would be, so to speak, invisible to natural selection; in that case it would be unlikely that the great preponderance of true belief over false required by reliability would be forthcoming.[11] So the question of the value of P(R/N&E) really turns on the relationship between belief and behavior. Our having evolved and survived makes it likely that our cognitive faculties are reliable and our beliefs are for the most part true, only if it would be impossible or unlikely that creatures more or less like us should behave in fitness-enhancing ways but nonetheless hold mostly false beliefs.[12]

Is this impossible or unlikely? That depends upon the relation between belief and behavior. What would or could that relation be? To try to guard against interspecific chauvinism, I suggested that we think, not about ourselves and our behavior, but about a population of creatures a lot like us on a planet a lot like earth (Darwin suggested we think about monkeys in this connection). These creatures are *rational*: that is, they form beliefs, reason, change beliefs, and the like. We imagine furthermore that they and their cognitive systems have evolved by way of the mechanisms to which contemporary evolutionary theory directs our attention, unguided by the hand of God or anyone else. Now what is P(R/N&E), specified, not to us, but to them? To answer, we must think about the relationship between their beliefs and their behavior. There are four mutually exclusive and jointly exhaustive possibilities.[13]

[11] Alternatively, we might say that the probability here is inscrutable, such that we can't make an estimate of it. Granted: it is unlikely that a large set of beliefs (comparable in size to the number of beliefs a human being has) should contain mainly truths; that gives us a reason for regarding the probability in question as low. On the other hand, we know something further about the relevant set of propositions, namely, that it is a set each member of which is believed by someone. How does this affect the probability? Perhaps we don't know what to say, and should conclude that the probability in question is inscrutable. (Here I am indebted to John Hare.)

[12] Must we concur with Donald Davidson, who thinks it is "impossible correctly to hold that anyone could be mostly wrong about how things are?" (See his "A Coherence Theory of Truth and Knowledge," in *Kant oder Hegel?* ed. Dieter Henrich [Stuttgart: Klett-Cotta Buchhandlung, 1983], 535.) No; what Davidson shows (if anything) is that it isn't possible for me to *understand* another creature unless I suppose that she holds mainly true beliefs. That may (or more likely, may not) be so; but it doesn't follow that there couldn't be creatures with mainly false beliefs, and a fortiori it doesn't follow that my own beliefs are mainly true. Davidson went on to argue that an *omniscient* interpreter would have to use the same methods we have to use and would therefore have to suppose her interlocutor held mostly true beliefs; given the omniscient interpreter's omniscience, he concluded that her interlocutor would in fact have mostly true beliefs. In so concluding, however, he apparently employs the premise that any proposition that would be believed by any omniscient being is true; this premise directly yields the conclusion that there is an omniscient being (since any omniscient being worth its salt will believe that there *is* an omniscient being), a conclusion to which Davidson may not wish to commit himself quite so directly. See *WPF*, 80–81.

[13] In *WPF* the argument involves five mutually exclusive and jointly exhaustive possibilities; here I telescope the first two.

(1) One possibility is *epiphenomenalism*:[14] their behavior is not caused by their beliefs. On this possibility, their movement and behavior would be caused by something or other—perhaps neural impulses—which would be caused by other organic conditions including sensory stimulation: but belief would not have a place in this causal chain leading to behavior. This view of the relation between behavior and belief (and other mental phenomena such as feeling, sensation, and desire) is currently rather popular, especially among those strongly influenced by biological science. The December 1992 issue of *Time* magazine reports that J. M. Smith, a well-known biologist, wrote "that he had never understood why organisms have feelings. After all, orthodox biologists believe that behavior, however complex, is governed entirely by biochemistry and that the attendant sensations—fear, pain, wonder, love—are just shadows cast by that biochemistry, not themselves vital to the organism's behavior." Smith could have added that (according to biological orthodoxy) the same goes for beliefs—at least if beliefs are not themselves just biochemical phenomena. If this way of thinking is right with respect to our hypothetical creatures, their beliefs would be *invisible* to evolution; and then the fact that their belief-forming mechanisms arose during their evolutionary history would confer little or no probability on the idea that their beliefs are mostly true, or mostly nearly true. Indeed, the probability of those beliefs' being for the most part true would have to be rated fairly low (or inscrutable). On N&E and this first possibility, therefore, the probability of R will be rather low.

(2) A second possibility is *semantic* epiphenomenalism: it could be that their beliefs do indeed have causal efficacy with respect to behavior, but not by virtue of their *content*. Put in currently fashionable jargon, this would be the suggestion that beliefs are indeed causally efficacious, but by virtue of their *syntax*, not by virtue of their *semantics*. On a naturalist or at least a materialist way of thinking, a belief could perhaps be something like a long-term pattern of neural activity, a long-term neuronal event. This event will have properties

[14] So called first by T. H. Huxley ("Darwin's bull dog"): "It may be assumed . . . that molecular changes in the brain are the causes of all the states of consciousness. . . . [But is] there any evidence that these states of consciousness may, conversely, cause . . . molecular changes [in the brain] which give rise to muscular motion? I see no such evidence. . . . [Consciousness appears] to be . . . completely without any power of modifying [the] working of the body, just as the steam whistle . . . of a locomotive engine is without influence upon its machinery." T. H. Huxley, "On the Hypothesis that Animals are Automata and its History" (1874), in *Method and Results* (London: Macmillan, 1893), 239–240. Later in the essay: "To the best of my judgment, the argumentation which applies to brutes holds equally good of men; and therefore, . . . all states of consciousness in us, as in them, are immediately caused by molecular changes of the brain-substance. It seems to me that in men, as in brutes, there is no proof that any state of consciousness is the cause of change in the motion of the matter of the organism. . . . We are conscious automata." (243–244). (Note the occurrence here of that widely endorsed form of argument, 'I know of no proof that not-*p*; therefore there is no proof that not-*p*; therefore *p*'.) However, I am here using the term to denote *any* view according to which belief isn't involved in the causal chain leading to behavior, whether or not that view involves the dualism that is apparently part of Huxley's version.

of at least two different kinds. On the one hand, there are its neurophysiological or electrochemical properties: the number of neurons involved in the belief, the connections between them, their firing thresholds, the rate and strength at which they fire, the way in which these change over time and in response to other neural activity, and so on. Call these *syntactical* properties of the belief. On the other hand, however, if the belief is really a *belief*, it will be the belief that *p* for some proposition *p*. Perhaps it is the belief that there once was a brewery where the Metropolitan Opera House now stands. This proposition, we might say, is the *content* of the belief in question. So in addition to its syntactical properties, a belief will also have *semantical* properties[15]—for example, the property of being the belief that there once was a brewery where the Metropolitan Opera House now stands. (Other semantical properties: *being true or false*, *entailing that there has been at least one brewery*, *being consistent with the proposition that all men are mortal*, and so on.) And this second possibility is that belief is indeed causally efficacious with respect to behavior, but by virtue of the *syntactic* properties of a belief, not its semantic properties. If the first possibility is widely popular among those influenced by biological science, this possibility is widely popular among contemporary philosophers of mind; indeed, Robert Cummins goes so far as to call it the "received view."[16]

On this view, as on the last, P(R/N&E) (specified to those creatures) will be low. The reason is that truth or falsehood are, of course, among the *semantic* properties of a belief, not its syntactic properties. But if the former aren't involved in the causal chain leading to behavior, then once again beliefs—or rather, their semantic properties, including truth and falsehood—will be invisible to natural selection.[17] But then it will be unlikely that their beliefs are mostly true, and hence unlikely that their cognitive faculties are reliable. The probability of R on N&E together with this possibility (as with the last), therefore, will be relatively low.

(3) It could be that beliefs are causally efficacious—'semantically' as well as

[15] Granted: the analogies between these properties and syntax and semantics is a bit distant and strained; here I am just following current custom.

[16] *Meaning and Mental Representation* (Cambridge: MIT Press, 1989), 130. In *Explaining Behavior* (Cambridge: MIT Press, 1988) Fred Dretske makes a valiant (but in my opinion unsuccessful) effort to explain how, given materialism about human beings, it could be that beliefs (and other representations) play a causal role in the production of behavior by virtue of their content or semantics. Part of the problem is that Dretske's account implies that there are no distinct but logically equivalent beliefs, and indeed no distinct but causally equivalent beliefs.

[17] We must also consider here the possibility that the syntax and semantics of belief are the effects of a common cause: perhaps there is a cause of a belief's having certain adaptive syntactic properties, which also causes the belief to have the semantic properties it does (it brings it about that the event in question is the belief that *p* for some proposition *p*); and perhaps this cause brings it about that a *true* proposition is associated with the belief (the neuronal event) in question. (Here I was instructed by William Ramsey and Patrick Kain.) What would be the likelihood, given N&E, that there is such a common cause at work? I suppose it would be relatively low: why should this common cause associate *true* propositions with these neuronal events? But perhaps the right answer is not that the probability in question is low, but that it is inscrutable.

'syntactically'—with respect to behavior, but *maladaptive*: from the point of view of fitness these creatures would be better off without them. The probability of R on N&E together with this possibility, as with the last two, would also seem to be relatively low.

(4) Finally, it could be that the beliefs of our hypothetical creatures are indeed both causally connected with their behavior and also adaptive. (I suppose this is the commonsense view of the connection between behavior and belief in our own case.) What is the probability (on this assumption together with N&E) that their cognitive faculties are reliable; and what is the probability that a belief produced by those faculties will be true? I argued that this probability isn't nearly as high as one is initially inclined to think. For one thing, if behavior is caused by *belief*, it is also caused by *desire* (and other factors—suspicion, doubt, approval and disapproval, fear—that we can here ignore). For any given adaptive action, there will be many belief-desire combinations that could produce that action; and very many of those belief-desire combinations will be such that the belief involved is false.

So suppose Paul is a prehistoric hominid; a hungry tiger approaches. Fleeing is perhaps the most appropriate behavior: I pointed out that this behavior could be produced by a large number of different belief-desire pairs. To quote myself:

> Perhaps Paul very much *likes* the idea of being eaten, but when he sees a tiger, always runs off looking for a better prospect, because he thinks it unlikely that the tiger he sees will eat him. This will get his body parts in the right place so far as survival is concerned, without involving much by way of true belief. . . . Or perhaps he thinks the tiger is a large, friendly, cuddly pussycat and wants to pet it; but he also believes that the best way to pet it is to run away from it. . . . or perhaps he thinks the tiger is a regularly recurring illusion, and, hoping to keep his weight down, has formed the resolution to run a mile at top speed whenever presented with such an illusion; or perhaps he thinks he is about to take part in a sixteen-hundred-meter race, wants to win, and believes the appearance of the tiger is the starting signal; or perhaps. . . . Clearly there are any number of belief-cum-desire systems that equally fit a given bit of behavior.[18]

Accordingly, there are many belief-desire combinations that will lead to the adaptive action; in many of these combinations, the beliefs are false. Without further knowledge of these creatures, therefore, we could hardly estimate the probability of R on N&E and this final possibility as high.

A problem with the argument as thus presented is this. It is easy to see, for just *one* of Paul's actions, that there are many different belief-desire combinations that yield it; it is less easy to see how it could be that most or all of his beliefs could be false but nonetheless adaptive or fitness enhancing. Could Paul's beliefs really be mainly false, but still lead to adaptive action? Yes in-

[18] *WPF*, 225–226.

deed; perhaps the simplest way to see how is by thinking of systematic ways in which his beliefs could be false but still adaptive. Perhaps Paul is a sort of early Leibnizian and thinks everything is conscious (and suppose that is false); furthermore, his ways of referring to things all involve definite descriptions that entail consciousness, so that all of his beliefs are of the form *That so-and-so conscious being is such-and-such*. Perhaps he is an animist and thinks everything is alive. Perhaps he thinks all the plants and animals in his vicinity are witches, and his ways of referring to them all involve definite descriptions entailing witchhood. But this would be entirely compatible with his belief's being adaptive; so it is clear, I think, that there would be many ways in which Paul's beliefs could be for the most part false, but adaptive nonetheless. From a naturalistic point of view, furthermore, we need not restrict ourselves to merely possible examples. Most of mankind has endorsed supernatural beliefs of one kind or another; according to the naturalist, such beliefs are adaptive though false.

What we have seen so far is that there are four mutually exclusive and jointly exhaustive possibilities with respect to that hypothetical population: epiphenomenalism simpliciter, semantic epiphenomenalism, the possibility that their beliefs are causally efficacious with respect to their behavior but maladaptive, and the possibility that their beliefs are both causally efficacious with respect to behavior and also adaptive. P(R/N&E) will be the weighted average of $P(R/N\&E\&P_i)$ for each of the four possibilities P_i—weighted by the probabilities, on N&E, of those possibilities. The probability calculus gives us a formula here:

$$P(R/N\&E) = (P(R/N\&E\&P_1) \times P(P_1/N\&E)) + (P(R/N\&E\&P_2) \times P(P_2/N\&E)) + (P(R/N\&E\&P_3) \times P(P_3/N\&E)) + (P(R/N\&E\&P_4) \times P(P_4/N\&E)).$$

Of course the very idea of a calculation (suggesting, as it does, the assignment of specific real numbers to these various probabilities) is laughable: the best we can do are vague estimates. Still, that will suffice for the argument. Now let's agree that P_3—the proposition that belief enters the causal chain leading to behavior both by virtue of neurophysiological properties and by virtue of semantic properties, but is nevertheless maladaptive—is very unlikely; it is then clear from the formula that its contribution to P(R/N&E) can safely be ignored. Note further that epiphenomenalism simpliciter and semantic epiphenomenalism unite in declaring or implying that the *content* of belief lacks causal efficacy with respect to behavior; the content of belief does not get involved in the causal chain leading to behavior. So we can reduce these two possibilities to one: the possibility that the content of belief has no causal efficacy. Call this possibility '–C'. What we have so far seen is that the probability of R on N&E&–C is low or inscrutable, and that the probability of R on N&E&C is also inscrutable or at best moderately high. We can therefore simplify (1) to

(2) P(R/N&E) = P(R/N&E&C) × P(C/N&E) + P(R/N&E&–C) × P(–C/N&E),

i.e., the probability of R on N&E is the weighted average of the probabilities of R on N&E&C and N&E&–C (weighted by the probabilities of C and –C on N&E).

We have already noted that the left-hand term of the first of the two products on the right side of the equality is either moderately high or inscrutable; the left-hand term of the second product is either low or inscrutable. What remains is to evaluate the weights, the right-hand terms of the two products. So what is the probability of –C, given N&E: what is the probability that one or the other of the two epiphenomenalistic scenarios is true? Note that according to Robert Cummins, semantic epiphenomenalism is in fact the received view as to the relation between belief and behavior.[19] That is because it is extremely hard, given materialism, to envisage a way in which the content of a belief could get causally involved in behavior. If a belief just is a neural structure of some kind—a structure that somehow possesses content—then it is exceedingly hard to see how content can get involved in the causal chain leading to behavior. For if a given such structure had had a different content but the same neurophysiological properties, its causal contribution to behavior, one thinks, would be the same. What causes the muscular contractions involved in behavior are physiological states of the nervous system, including physiological properties of those structures that constitute beliefs; the content of those beliefs appears to be causally irrelevant. So it is exceedingly hard to see, given N&E, how the content of a belief can have causal efficacy.

It is exceedingly hard to see, that is, how epiphenomenalism—semantic or simpliciter—can be avoided, given N&E. (There have been some valiant efforts but things don't look hopeful.) So it looks as if P(–C/N&E) will have to be estimated as relatively high; let's say (for definiteness) .7, in which case P(C/N&E) will be .3. Let's also estimate that P(R/N&E&–C) is, say, .2. Then P(R/N&E) will be at most .45, less than ½. Of course we could easily be wrong; the argument for a low estimate of P(R/N&E) is by no means irresistible; our estimates of the various probabilities involved in estimating P(R/N&E) with respect to that hypothetical population were (naturally enough) both imprecise and poorly grounded. You might reasonably hold, therefore, that the right course here is simple agnosticism: one just doesn't know what P(R/N&E) is. You doubt that it is very high; but you aren't prepared to say that it is low: you have no definite opinion at all as to what that probability might be. Then this probability is *inscrutable* for you. This too seems a sensible attitude to take. The sensible thing to think, then, is that P(R/N&E) is either low or inscrutable.

Now return to Darwin's Doubt, and observe that if this is the sensible attitude to take to P(R/N&E) specified to that hypothetical population, then it

[19] *Meaning and Mental Representation*, 130.

will also be the sensible attitude toward P(R/N&E) specified to us. If N&E is true with respect to us, then we are relevantly like them in that our cognitive faculties have the same kind of origin and provenance as theirs are hypothesized to have. And the next step in the argument was to point out that each of these attitudes—the view that P(R/N&E) is low and the view that this probability is inscrutable—gives the naturalist-evolutionist a defeater for R. It gives him a reason to doubt it, a reason not to affirm it. I argued this by analogy. Among the crucially important facts, with respect to the question of the reliability of a group of cognitive faculties, are facts about their origin. Suppose I believe that I have been created by an evil Cartesian demon who takes delight in fashioning creatures who have mainly false beliefs (but think of themselves as paradigms of cognitive excellence): then I have a defeater for my natural belief that my faculties are reliable. Turn instead to the contemporary version of this scenario, and suppose I come to believe that I have been captured by Alpha-Centaurian superscientists who have made me the subject of a cognitive experiment in which I have been given mostly false beliefs: then, again, I have a defeater for R. But to have a defeater for R it isn't necessary that I believe that in fact I have been created by a Cartesian demon or been captured by those Alpha-Centaurian superscientists. It suffices for me to have such a defeater if I have considered those scenarios, and the probability that one of those scenarios is true, is inscrutable for me. It suffices if I have considered those scenarios, and for all I know or believe one of them is true. In these cases too I have a reason for doubting, a reason for withholding[20] my natural belief that my cognitive faculties are in fact reliable.

Now of course defeaters can be themselves defeated. For example, I know that you are a lifeguard and believe on that ground that you are an excellent swimmer. Then I learn that 45 percent of Frisian lifeguards are poor swimmers, and I know that you are Frisian: this gives me a defeater for the belief that you are a fine swimmer. Then I learn still further that you have graduated from the Department of Lifeguarding at the University of Leeuwarden and that one of the requirements for graduation is being an excellent swimmer: that gives me a defeater for the defeater of my original belief: a defeater-defeater as we might put it.[21] But (to return to our argument) can the defeater the naturalist has for R be in turn defeated? I argued that it can't.[22] It could be defeated only by something—an argument, for example—that involves some other belief (perhaps as a premise). Any such belief, however, will be subject to the very same defeater as R is. So this defeater can't be defeated.[23] But if I have an undefeated defeater for R, then by the same token I have an unde-

[20] I shall use this term to mean *failing to believe*, so that I withhold *p* if either I believe its denial or I believe neither it nor its denial.

[21] As in fact John Pollock *does* put it; see his *Contemporary Theories of Knowledge* (Totowa, N.J.: 1986), 38–39.

[22] *WPF*, 233–234.

[23] I'll qualify this when we get to the subject of loops.

feated defeater for any other belief B that my cognitive faculties produce, a reason to be doubtful of that belief, a reason to withhold it. For any such belief will be produced by cognitive faculties that I cannot rationally believe to be reliable. Clearly the same will be true for any belief they produce: if I can't rationally believe that the faculties that produce that belief are reliable, I have a reason for rejecting the belief. So the devotee of N&E has a defeater for just any belief he holds—a defeater, as I put it, that is ultimately undefeated. This means, then, that he has an ultimately undefeated defeater for N&E itself. And that means that the conjunction of naturalism with evolution is self-defeating, such that one can't rationally accept it. I went on to add that anyone who accepts naturalism ought also to accept evolution; evolution is the only game in town, for the naturalist, with respect to the question of how all this variety of flora and fauna has arisen. If that is so, finally, then naturalism simpliciter is self-defeating and cannot rationally be accepted—at any rate by someone who is apprised of this argument and sees the connections between N&E and R.

PART I

SCIENCE AND EVOLUTION

Naturalism Defended

William Ramsey

I. Introduction

Alvin Plantinga has recently put forward an ingenious argument designed to show that naturalistic accounts of our cognitive mechanisms are self-refuting.[1] Very roughly, Plantinga argues that if one assumes that a given cognitive system is the product of naturalistic evolutionary processes (i.e., no guidance from a higher being), then one should also believe that it is probably unreliable (or at least that its reliability is uncertain). Since a naturalist thinks our own cognitive faculties came about this way, a naturalist has little reason to trust our faculties and, consequently, any belief they might produce. But since commitment to naturalism is itself just such a belief, naturalism gives the naturalist a reason for doubting naturalism (it is, as Plantinga puts it, a "self-defeating" position).

It would be a bit of an understatement to say that there are many avenues that could be (and have been) adopted by the naturalist in responding to Plantinga's argument. In this chapter I'm going to focus on only one specific element of his position. A critical premise of Plantinga's argument is the claim that evolutionary theory, by itself, gives us little reason for supposing that our beliefs are, for the most part, true. In other words, according to

Versions of this essay were presented at the World Congress on Philosophy (Boston, 1998) and at the University of Nevada at Las Vegas (March 1999). Comments from these audiences were quite helpful. I'm also grateful to Michael Bergmann, Stephen Downes, Ernan McMullin, Richard Otte, Michael Rea, Leopold Stubenberg, Todd Jones, and, of course, Alvin Plantinga for a number of helpful discussions and suggestions.

[1] Plantinga, *Warrant and Proper Function* (New York: Oxford University Press, 1993); and "Naturalism Defeated" (unpublished).

Plantinga, there are good reasons for thinking that the position I'll call "evolutionary reliabilism"—the view that natural selection tends to favor reliable belief-producing mechanisms—is mistaken. If Plantinga is right about this—if the probability is fairly high that evolutionary processes would follow a course where unreliable mechanisms are selected—then we cannot have much assurance that, in fact, evolution didn't follow such a course.

The argument against evolutionary reliabilism is a key component of Plantinga's larger argument against naturalism. If it should turn out that evolutionary reliabilism survives his attack, then Plantinga's case against naturalism cannot get off the ground. The aim of this chapter is to convince you that, in fact, Plantinga's attempts to undermine evolutionary reliabilism do not succeed. I'll begin by demonstrating how evolutionary reliabilism has a fair degree of initial plausibility for the naturalist. This will require discussing an important issue Plantinga himself raises; namely, the link between truth and adaptive behavior. After explicating the proper way to understand this matter, I'll then show why Plantinga's two main arguments against evolutionary reliabilism fail to pose a serious challenge.

II. The Adaptive Role of Truth

What contribution does truthful thinking make to adaptive behavior? Plantinga himself suggests that naturalism allows the truthfulness of cognitive states to be related (or unrelated) to behavior in a variety of different ways, but that only one of these has the possibility of generating adaptive behavior. However, the relation between truth and behavior is more complex than Plantinga suggests, and he overlooks the most sensible way for a naturalist (or anyone) to regard the connection between the two. If we are going to appreciate the plausibility of evolutionary reliabilism, we need first to get clear on how truth can be a property of beliefs that bestows an advantage on cognitive systems.

While there is considerable disagreement among naturalists about the ontological status of some of the semantic properties of beliefs (i.e., disagreement about whether content ultimately is or is not "in the head"), there is much greater agreement about the status of the semantic properties we care about here, namely, truth and falsehood. Most naturalistic philosophers of mind—indeed, I think just about anyone who considers the issue carefully—would agree that the truth value of a belief is a relational property that depends, in part, on what those beliefs are about. In other words, a belief's truth value does not supervene on the belief's intrinsic/neurophysiological properties; truthfulness can vary while the neurophysiological properties remain the same. Hence, a belief's truth value, though partly dependent upon what is going on inside the head, is not reducible to neurological matters.

This irreducibility appears to raise a serious difficulty for anyone who wants to claim that a belief's truth value is relevant to evolution and adaptive behav-

ior. Consider the following argument: What takes place inside the head is what ultimately matters for the production of behavior; that is, a belief's causal role in the production of behavior arises entirely from its intrinsic, physical properties. If you want to know why a mental state does what it does, then you need only look at its physical (neurochemical or "syntactic") properties.[2] Since the truth value of a belief is not reducible to those physical properties, then the truth value is irrelevant to the belief's causal role. But if a belief's truth value is irrelevant to its causal role, then it must be irrelevant to the production of behavior. And if it is irrelevant to the production of behavior, then it must be irrelevant to the production of *adaptive* behavior. Since it is adaptive behavior that really matters for natural selection, truth would then be a property of beliefs that makes no contribution to survival or natural selection. As Plantinga himself puts it, "if [truth and falsehood] aren't involved in the causal chain leading to behavior, then once more beliefs—or rather, their semantic properties, including truth and falsehood—will be invisible to natural selection."[3]

Although at first glance this argument seems plausible, it begins to lose that plausibility when we consider the various different ways in which relational properties can come to be causally salient. As it turns out, in selectional phenomena of precisely the sort that matter for evolution, properties can be causally relevant in ways that the above argument overlooks. Perhaps the easiest way to see this is with an example. Suppose ten individuals are given ten different maps, only of one which provides an accurate route to buried treasure. If we want to explain the behavior of any given individual (including the one with the right map) then all we need to appeal to are the actual directions—the instructions and arrows and such—physically displayed on that individual's map. The intrinsic features of the map are the causally relevant properties that produce the behavior in question, and the accuracy or inaccuracy of any given map would be, by and large, irrelevant. But now suppose that instead of explaining the actual behaviors of the treasure hunters, we want to know why the behavior of one hunter is of a certain sort—namely, of the sort that actually succeeds in finding the treasure. Clearly then it would be appropriate to appeal to the further fact that the lucky hunter's map is the accurate one. That is, if we want to explain how one hunter actually manages to find the treasure, the accuracy of his map becomes *the* causally salient feature—it is what makes it the case that this particular hunter succeeds while the others fail. Indeed, if finding the treasure was in some way critical for survival, it would be the feature of the map that makes it the case that this hunter survives while others do not.

For those familiar with evolutionary explanations, it is not surprising that a relational property that fails to supervene on intrinsic physical features can,

[2] Something like this once prompted Jerry Fodor and others to endorse a sort of "methodological solipsism" in the cognitive sciences (see Fodor, "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology," in *Representations* [Cambridge: MIT Press, 1981], 225–253).

[3] "Naturalism Defeated," 8.

nonetheless, be a causally salient property that brings about a certain state of affairs. In natural selectional processes, irreducible relational properties often play a large role in determining which members of a set fare the best. For example, survival sometimes depends on an organism's ability to hide in a specific environment. Having a good camouflage is often a key adaptive feature that is favored by natural selection. However, since the quality of a given camouflage depends, in part, on the environment, a good camouflage does not supervene on an organism's intrinsic physical makeup. This is a fairly common case where a non-supervening property (having a good camouflage) enhances a creature's reproductive fitness and, hence, becomes a property for which there is considerable selection pressure. The irreducibility of a good camouflage in no way undermines its central importance to natural selection.[4]

What all of this suggests is that there is a perfectly plausible way for truth to be a property that, although irreducible, is nevertheless favored by a process of natural selection. Indeed, since some naturalists characterize beliefs as the "maps by which we steer,"[5] a slight modification of our earlier example shows us exactly how this would work. Suppose we have ten creatures competing for a scarce resource such as food, and suppose that only one of these creatures (Bob) possesses accurate beliefs concerning the whereabouts of the food. If all we want is an explanation of Bob's immediate motor behavior, then we needn't appeal to the truth or falsehood of any of his beliefs. But if we want to know why Bob's behavior proves successful while his cohorts die out, then it clearly does matter that his behavior is generated by true beliefs. In a narrow and somewhat mechanistic sense, the truth value of a belief is indeed epiphenomenal, since it can vary without effecting the causal powers of the cognitive mechanisms that generate behavior. But in another broader sense, it clearly is *not* epiphenomenal, since variance in a belief's truth value can alter the broader *consequences* of the behavior, making it either successful or unsuccessful (from an evolutionary perspective). To adapt a recent slogan: truth doesn't make the behavior, it makes the behavior better. Hence, truth and reliability are exactly the sort of features for which there can indeed be considerable selection pressure.

So the naturalist has a very plausible and compelling account of the relation between beliefs and behavior which makes truth (1) a causally relevant feature of beliefs and (2) causally relevant in a way that enhances reproductive fitness. If we add to this the further claim that a preponderance of true beliefs is, in fact, something that our cognitive mechanisms have been selected to

[4] See also Eliot Sober, *The Nature of Selection* (Cambridge: MIT Press, 1984). This sort of causality (between non-supervening relational properties) also occurs outside of biology. If a something has the relational property of being the heaviest particle in a solution, this will often cause it to acquire the other relational property of being the lowest particle (relative to other particles). Here it seems clear that one relational property brings about another. I'm grateful to Michael Rea for discussions on this point.

[5] F. P. Ramsey, *The Foundations of Mathematics, and Other Logical Essays* (London: Routledge and Kegan Paul, 1931); F. Dretske, *Explaining Behavior* (Cambridge: MIT Press/Bradford), 1988.

produce, then we have the basic idea behind evolutionary reliabilism. Putting things very roughly (and, no doubt, too simplistically), evolutionary reliabilism claims that cognitive mechanisms come to be adaptive, in part, by generating desires that correspond with reproductive fitness (e.g., a desire to reproduce, a desire to stay alive, etc.) and accurate representations that enable the organism to get around and satisfy those desires. On this view, the truthfulness of our beliefs actually helps *explain* their adaptiveness. While belief-producing mechanisms cannot be altered to produce adaptive cognitive states *directly*, they can be altered to produce more accurate cognitive states. The accuracy of those states, in turn, makes them more inclined to generate fitness-enhancing behavior.

Why believe evolutionary reliabilism? While this is a huge and complex topic, one quick answer is that there does not appear to be a serious alternative that explains how our cognitive states, and hence our cognitive faculties, come to be adaptive. It is, in fact, extremely difficult to give a serious adaptational account of cognitive processes in which truth (or something close to it, like accuracy) does not matter for mental representations. This is not to say that truth is the *only* property of beliefs that matters—no doubt a wide range of other properties are highly important as well (such as ease of processing, relevance, etc.). But these properties are, for the most part, beneficial only when conjoined with some degree of verisimilitude. While there are a few accounts of natural selection producing very limited and localized forms of unreliabilism, one of which will be discussed below, there appears to be no general theory of our cognitive faculties that demonstrates how natural selection could favor cognitive mechanisms that systematically generate false beliefs in a global way.

In his case against naturalism, Plantinga insists that this last claim is mistaken and that evolutionary reliabilism is not as plausible as I've tried to make it seem. To show this, he offers two ways in which our basic cognitive systems could be adaptive while generating falsehoods. Let's now look closely at these accounts.

III. Evolutionary Reliabilism Defended

The first sort of scenario Plantinga offers to demonstrate the adaptiveness of false thinking is one where the desires and goals of an organism are properly adjusted so that erroneous beliefs nevertheless produce adaptive behavior. The second scenario he offers is one where deep misconceptions that are claimed to falsify a large set of an organism's noetic states do so in a benign way that does not lead to maladaptive behavior. What I plan to demonstrate is that these two sorts of scenarios fail to pose a serious challenge to evolutionary reliabilism. The problem with the first scenario is that there is no good reason to think that the requisite desire/goal modification could come

about through a mechanism designed by evolutionary processes. The problem with the second scenario is that it fails to motivate the view that cognitive states could be systematically false in a way that challenges reliabilism.

A. Adjusting Goals and Desires

The first way Plantinga suggests natural selection could give rise to unreliable cognitive mechanisms is by generating suitable desires and goals that, when conjoined with appropriate false beliefs, give rise to fitness-enhancing behavior. As Plantinga puts it:

> For any adaptive action, there will be many belief-desire combinations that could produce that action; and very many of those belief-desire combinations will be such that the belief involved is false. . . . Perhaps Paul very much *likes* the idea of being eaten, but when he sees a tiger, always runs off looking for a better prospect, because he thinks it unlikely that the tiger he sees will eat him. This will get his body parts in the right place so far as survival is concerned, without involving much by way of true belief. . . . Or perhaps he thinks the tiger is a large, friendly, cuddly pussycat and wants to pet it; but he also believes that the best way to pet it is to run away from it.[6]

Plantinga's central point is quite clear. Once we have determined an appropriate course of action for an individual creature in a particular situation, it is always possible to imagine conjunctions of false beliefs with certain desires that would generate the adaptive behavior in question. In fact, it would seem that there is a huge array of possible false belief-desire pairs that for any given scenario would generate adaptive behavior.

By itself, however, this point establishes very little. The main point of contention is not whether there could exist false beliefs that, when coupled with certain desires, would produce adaptive behavior in certain conditions. No one—including the evolutionary reliabilist—would deny that. Rather, the critical question is whether or not a process of natural selection is likely to generate cognitive mechanisms that are systematically unreliable but somehow prove adaptive. Instead of giving us specific belief-desire pairs that produce life-preserving behavior in certain odd circumstances, Plantinga needs to describe general *mechanisms* or *processes* that would prove adaptive while giving rise to a preponderance of false beliefs. In other words, what Plantinga needs to explain in order to challenge evolutionary reliabilism is a mechanism or process that (a) could come about through evolution, and (b) generates mostly false beliefs over time, and (c) nevertheless proves adaptive. Plantinga's examples, though interesting, don't provide us with anything like this. They are simply unrelated scenarios in which erroneous thinking happens to prove lucky for the cognitive agent. What we need is a general account of

[6] "Naturalism Defeated," 9.

how an unreliable cognitive system can be *systematically* adaptive; what we are given instead are individual examples of clearly maladaptive systems fortuitously producing appropriate behavior.

I say "clearly maladaptive systems" because it is hard to see how a cognitive system that generates the belief that tigers are large, cuddly pussycats or the belief that the best way to get near something is to run away from it, to use Plantinga's examples, would prove adaptive in the long run. While errors of this sort may not get you killed in certain specific circumstances, a general disposition to generate similar sorts of errors will, down the road, get you into trouble. If your cognitive system is prone to those sorts of errors, then you aren't going to be around for long. So without some general account of unreliable cognitive processes or mechanisms that would nevertheless prove adaptive over time, there is no reason to view these isolated examples as a motivation for giving up evolutionary reliabilism.

Notice that the situation for the evolutionary reliabilist is very different, since the evolutionary reliabilist *does* provide us with a very clear and plausible story about how our cognitive mechanisms are *systematically* adaptive. Since it is clear how accurate representations tend to help an organism satisfy its desires and goals (in a way that inaccurate representations do not), evolutionary reliabilism gives us a mechanism that makes the value of these representations quite clear. For those like Plantinga who reject evolutionary reliabilism but nevertheless maintain that beliefs somehow give rise to fitness-promoting behavior, there is a burden to provide an alternative explanation of how that happens. There is a need to describe some other mechanism or some other feature of beliefs, besides truth, that would systematically generate the right sort of behavior. If there is no such account, then we have no reason for thinking evolutionary processes would produce unreliable cognitive mechanisms.

At this point it might be asked why we can't simply appeal to adaptiveness itself as the special feature that all of the false belief-desire pairs have in common. If they are, in fact adaptive, doesn't that give us an alternative feature, besides truth, that can explain the adaptive value of beliefs? The problem with the response is that simple adaptiveness is not a basic feature that evolutionary processes can bring about directly, or that cognitive mechanisms or cognitive states can have in any simple or irreducible sense. Insofar as a given cognitive system is adaptive, or generates adaptive behavior, it does so only by virtue of some other feature or quality that it possesses. In fact, no aspect of an organism is adaptive in any sort of basic or primitive sense. Defense mechanisms, reproductive systems, foraging strategies, and the like get to be adaptive only by being *something else* first (like deceptive, or highly selective). If we know a given cognitive system is adaptive, the next question to ask is how, exactly, it gets to be that way. According to the evolutionary reliabilist, a cognitive system gets to be that way through modifications that produce more accurate representations. Adaptiveness is the explanandum that the reliabilist

picture and whatever alternative must explain. The problem with Plantinga's first reason for rejecting evolutionary reliabilism is that he gives us no other alternative.

Another possible response would be to suggest that a cognitive mechanism could employ a strategy similar to the one Plantinga himself uses in contriving false belief-desire pairs; namely, by seeing what needs to be done and cooking up the odd belief-desire pairs that lead to fitness-promoting behavior. However, even if we allowed the problematic homuncularism implied by such a strategy, the obvious difficulty with this proposal is that Plantinga himself does something that an unreliable system couldn't do. Plantinga is able to construct such false belief-desire pairs because he has accurate *knowledge* about various circumstances and proper behavior. For instance, Plantinga uses his own true beliefs about tigers to figure out what one ought to do when confronted with one; then, using this, he concocts weird false belief-desire pairs that will nonetheless generate appropriate behavior. However, such a strategy clearly wouldn't work for an *un*reliable cognitive system, which *ex hypothesi* doesn't have access to a store of knowledge about tigers, or much of anything else.

A third response might be to deny that the relevant units of selection are general purpose belief-producing mechanisms, and insist instead that individual false beliefs (or at least dispositions to produce specific false belief-desire pairs) are the things for which there is selection pressure. That way, natural selection itself could provide the mechanism that explains a preponderance of helpful false belief-desire pairs. Organisms with a genetic predisposition to think certain specific combinations of false beliefs and desires (like the false belief that to approach a tiger you must run away, conjoined with the desire to pet a tiger) would be selected over those with less happy combinations. However, this proposal implies an extreme sort of nativism that most naturalists would reject. For this scenario to work, we must assume beliefs with specific content are in some way genetically encoded. By contrast, the standard naturalistic picture of how cognition evolved would treat cognitive *mechanisms* and inference *strategies* as the sorts of things that are passed on from one generation to the next. Specific beliefs are the *output* of those inherited capacities, dependent upon a myriad of factors beyond genetic inheritance (such as learning).[7] If Plantinga's account is to have any plausibility from a naturalistic perspective, we must assume that selection pressure is directed toward cognitive mechanisms, not individual beliefs.

Finally, it might be asked why such an explanation or account is necessary? Why couldn't Plantinga just posit a desire-forming mechanism that, when connected with an unreliable belief-forming mechanism, always spontaneously generates unrelated desires of the right sort? Or why can't Plantinga

[7] I'm grateful to Leopold Stubenberg for pointing out this nativist picture.

simply stipulate that the agent's belief-forming mechanism always happens to yield just the right sort of false belief that, in conjunction with whatever wacky desire the agent has at the time, always preserves fitness?

The obvious problem with this proposal is that such a mechanism has absolutely no plausibility from the standpoint of naturalism. Indeed, from any standpoint the operation of such a mechanism would be utterly mysterious—inexplicably generating unrelated false belief-desire pairs that just happen to always keep the agent out of trouble. While such a mechanism may be conceivable, and would indeed prove adaptive, there is no reason to think a process of natural selection could ever construct such a device. Put bluntly, the operation of such a mechanism would be miraculous, and naturalism doesn't allow for miraculous mechanisms.

What might a nonmiraculous mechanism look like, one that could challenge evolutionary reliabilism? As noted above, constructing a plausible account of such a mechanism or process is extremely difficult. One possibility would be to appeal to some other general feature of beliefs besides truth and explain how it could allow false beliefs to be adaptive. One such possible feature, suggested by Richard Feldman, is optimism.[8] We often hear about the power of positive thinking, that successful people are unusually optimistic. Perhaps an account could be given in which an optimistic but unreliable belief-producing mechanism can be seen as more conducive toward survival than a reliable mechanism—where the end result would be a happy but mostly mistaken cognitive agent.

Yet it is difficult to see how, in the long run, such an account could be made to work. It is hard to see how an agent who regularly discounts dangers, overestimates his or her own abilities, and carries a number of false expectations will do better than an agent with a more realistic picture of things. Positive thinking generally works when it is coupled with *accurate* thinking. It seems the only environment in which extremely optimistic reasoning would prove successful is one that is indeed utopian—where there are very few dangers and many advantages. But in such an environment, an optimistic mechanism would also be a *reliable* one. What would lead to the agent's success in such a setting is not that the beliefs are optimistic but that they are, as it turns out, correct.

Going the other way, perhaps a case can be made for thinking that pessimism is an adaptive feature of beliefs that could trump truth.[9] In fact, many have suggested that the case for positive thinking and high self-esteem has been overstated.[10] Indeed, there are many successful people who often ex-

[8] Feldman, "Rationality, Reliability, and Natural Selection," *Philosophy of Science* 55 (1988): 226.

[9] Ibid.

[10] See, for example, R. Dawes, *House of Cards: Psychology and Psychotherapy Built on Myth* (New York: Free Press, 1994).

press negative thoughts about their circumstances and abilities. So maybe the general property of being mistakenly negative could replace truthfulness as a fitness-promoting feature of beliefs.

Once again, however, it is tough to see how this proposal could be made plausible. With one possible exception, to be discussed below, it would appear that mistakenly negative thoughts are generally less apt to promote fitness than true thoughts. If you mistakenly think you can't improve your circumstances, then you will fail to attempt to do so; if you mistakenly think that predators are everywhere, then you will fail to seek things you need; if you mistakenly think most foods are poisonous, then you will wind up malnourished, and so on. If a wrong-headed pessimist is competing for resources with a right-headed realist, there seems little chance that the pessimist will, in the long run, win out.

There is, however, one sort of scenario in which it seems organisms with an overly cautious belief-forming mechanism could be selected over those with a more reliable one. Since the specific details of the scenario, sometimes referred to as the Garcia effect, have already been discussed at length in the literature, I'll simply spell out the core thesis here.[11] The basic idea behind the Garcia effect is that in certain cases the cost of falsehood is far greater for some beliefs than it is for others; hence to insure veracity regarding the few beliefs that really matter, an inference strategy is selected where a large number of falsehoods are allowed. For example, suppose you live in an environment in which there are occasional encounters with a nasty predator, and frequent encounters with an organism that resembles the predator but is actually harmless. Also suppose that the behavior you adopt for avoiding the predator is relatively easy (like simply ducking down a hole). Since the cost of false negatives (thinking the predator isn't around when it actually is) is so much higher than the cost of false positives (thinking the predator *is* around when it actually isn't), then the most important strategy is one that avoids false negatives. One such strategy is to assume that everything that looks like the predator actually *is* the predator. Such a strategy would not be very reliable, since more often than not it would generate the belief that a predator is nearby when one isn't. It would, however, be superior to a more reliable strategy that allowed fewer false positives but a few more false negatives. This seems to provide us with one way that natural selection could favor what might be viewed as an unreliable belief-forming strategy over a more reliable one.

Yet it is far from clear that this sort of scenario actually undermines a sophisticated version of evolutionary reliabilism. What the Garcia effect tells us is that the simple percentage of true beliefs is not the sole consideration that

[11] See J. Garcia et al., "Biological Constraints on Conditioning," in *Classical Conditioning: Current Research and Theory*, vol. 2, ed. A. Black and W. Prokasy (New York: Appleton-Century Crofts, 1972); Stephen Stich, *The Fragmentation of Reason* (Cambridge: MIT Press, 1990), 61–63; E. Stein, *Without Good Reason* (Oxford: Clarendon Press, 1996), 190–197.

comes into play in determining the advantage bestowed by a given cognitive system. But no sophisticated reliabilist would ever deny this. Notice, for example, that no sophisticated evolutionary reliabilist would claim that a cognitive mechanism that generates only true mathematical beliefs would be better than a more fallible system that generates beliefs about the environment. The content of the beliefs obviously matters a great deal. It is the truth ratio of certain *types of* beliefs that matters. And just as it can be more important to be right when you hold a belief with the content "the environment is safe" than when you hold one with the content "$25 \times 3 = 75$", so too it can be more important than when you hold one with the content "the environment is dangerous". So the Garcia effect is really just a special case of something a sensible reliabilist would have admitted all along. The Garcia effect still treats truth as a critical property of beliefs—it just reminds us that it is more important to get some things right rather than others. Indeed, one could say that mechanisms of this sort *are* selected to produce a higher ratio of true beliefs with a certain content over false beliefs with the same content—so in one sense they actually *are* reliable after all.

Even if we dismiss this somewhat Pickwickian view of a reliable mechanism, the Garcia effect would fail to provide much support to Plantinga unless it could be shown that it pertains to enough of our cognitive mechanisms and strategies to call into question our beliefs about naturalism. At the very least, we need something like an argument showing that the truthfulness of scientific beliefs needs to be sacrificed to insure the truthfulness of beliefs more directly linked to reproductive fitness. Since many of the inference strategies used for the former are the same ones used for the latter (like induction), it is very hard to see how this could be done.[12] But without this sort of argument, the Garcia effect does very little to help Plantinga's argument against naturalism.

B. Deep Misconceptions

Plantinga appears to be aware of some of the difficulties raised in the last section, since he admits, "It is easy to see, for just *one* of Paul's actions, that there are many different belief-desire combinations that yield it; it is less easy to see how it could be that most of all of Paul's beliefs could be false but nonetheless adaptive or fitness enhancing."[13] To help himself out of this dilemma, Plantinga introduces a second way in which an organism's beliefs might be adaptive but mostly false. He supposes that the organism suffers from some sort of profound misconception that undermines the veracity of its beliefs without undermining the beliefs' usefulness. Returning to the hypothetical Paul, Plantinga suggests,

[12] Elliott Sober, "The Evolution of Rationality," *Synthese* 46 (1981): 96–120.
[13] "Naturalism Defeated," 10.

> Perhaps Paul is a sort of early Leibnizian and thinks everything is conscious (and suppose that is false); furthermore, his way of referring to things all involve definite descriptions that entail consciousness, so that all of his beliefs are of the form *that so-and-so conscious being is such-and-such*. Perhaps he is an animist and thinks everything is alive . . . But this would be entirely compatible with his beliefs being adaptive; so it is clear, I think, that there would be many ways in which Paul's beliefs could be for the most part false, but adaptive.

But is this as clear as Plantinga suggests? First, the argument Plantinga presents is incomplete, as it is not obvious how we go from the claim that Paul is a pan-psychist or an animist to the claim that most of his beliefs are false. The missing premise appears to be something like the following principle:

> P1: If any of your beliefs about some subject are false, then all (or most) of your beliefs about that subject are false.

But P1 is clearly false. No doubt some of the things I believe about a lot of topics are false; but it would be a mistake to conclude that all of my beliefs about those things are false. So P1 can't be the principle Plantinga would use to complete his argument. In fact, Plantinga's reference to the definite descriptions suggest that it isn't the falsehood of any old beliefs he has in mind, but only the falsehood of those that are used to pick out or refer to the subject in question. Yet, again, this needs clarification. Just because someone thinks that trees are conscious, and thinks of them under the description "those conscious trees", it clearly doesn't follow that they can't also think, truthfully, that trees grow in the ground, that they need sunlight to live, that some lose their leaves in the fall, and so on. What appears to be motivating Plantinga is the assumption that if someone suffers from some sort of *deep* misconception, and that misconception is employed in a belief, then the belief itself must be false. This suggests the following revised principle:

> P2: If you suffer from a deep misconception of some subject, then all (or most) of your beliefs about that subject are false.

By "deep misconception", I mean false beliefs about what might be considered essential or definitional properties, as opposed to merely contingent features. If P2 is correct, then it indeed seems possible for cognitive beings to have mostly false beliefs about things in their environment, and yet behave in a manner that is largely indistinguishable from creatures with true beliefs. But is P2 correct, or perhaps more important, is it a principle most naturalists would endorse?

I believe not. Consider first a sentence that might be uttered by someone, such as "Fire can be dangerous." Following the work of Saul Kripke and oth-

ers, there is now widespread agreement that in such an utterance the reference of the term "fire" is due in large measure to causal connections that exist between the speaker and actual instances of fire.[14] An important consequence of this view is that the speaker can refer to fire, and can predicate truths about fire, even if she suffers from profound ignorance and/or false beliefs about the nature of fire itself. Indeed, it is precisely this consequence that makes the causal theory of reference so appealing over competitors like the description theory. It seems quite obvious that a person can truthfully say fire is dangerous—and really be referring to fire—even though the person mistakenly thinks that fire is, say, essentially a substance.

So much for utterances, but what about beliefs? Well, if we adopt the very popular and naturalistically appealing Fodorian view that beliefs are complex representations made up of individual concepts, then what we just said about sentences of public language can also be said (and has been said) about sentences of a language of thought. If I believe "Fire is dangerous", then my belief, just like my utterance, is about whatever it is my fire concept denotes. What my fire concept denotes is largely determined by causal connections between my fire concept and fire. Hence, my fitness-enhancing belief that fire is dangerous is true, despite whatever other beliefs I have about fire. In fact, even if I explicitly think of fire every time under the description of "that unusual substance", I can still truthfully assign properties to fire. As Keith Donnellan has demonstrated, definite descriptions serve a referential as well as attributive function, just like names or predicates.[15]

It seems then, that Plantinga's second reason for claiming that a belief system can be (a) adaptive and (b) mostly false is not very compelling, at least not for a naturalist persuaded by popular theories of reference. Once again the problem is that the sort of error Plantinga discusses fails to lead to any broader sort of systematic falsehood that could be used to undermine reliabilism. Plantinga fails to give the naturalist a good reason for supposing that evolutionary reliabilism is any less probable than what she might have initially assumed.

IV. Conclusion

As I said in the introduction it is essential to Plantinga's argument against naturalism that he drive a wedge between the adaptiveness of belief-generating mechanisms and their reliability. As I've just shown, it seems that he has failed in this endeavor; consequently, he has failed in his larger endeavor to

[14] Kripke, *Naming and Necessity* (Cambridge: Harvard University Press, 1980).

[15] K. Donnellan, "Reference and Definite Descriptions," *Philosophical Review* 75 (1966): 281–304.

show that naturalism, conjoined with evolutionary epistemology, is self-refuting. In closing, I want to briefly discuss a couple of important issues that have been overlooked in Plantinga's discussion of reliabilism and evolution.

The first issue concerns the sort of unreliabilism Plantinga assumes could come about through a process of natural selection. There are, after all, many different ways our cognitive systems and strategies can go astray, and it makes sense to ask which of these a natural process is likely to produce. To conform with our everyday experience (in which we seem to enjoy a considerable degree of predictive and explanatory success), Plantinga's argument presupposes a very devious type of unreliabilism that is difficult to detect—where somehow our faculties conspire to give rise not just to error, but to full-blown duplicity of the sort that would make a Cartesian demon proud. On this model, our cognitive capacities are not only unreliable but downright deceiving in a systematic way.

But what is the prima facie likelihood that evolution would generate mechanisms of this sort? Far lower, it would seem, than the likelihood that it would simply generate reliable mechanisms. After all, this sort of deception would not be easy to pull off, and the mechanisms involved would need to be more sophisticated than those that are straightforwardly truth-tropic (that is why Descartes's demon was characterized as godlike). For example, it seems evolution would need to generate additional mechanisms that somehow ensured the divergent faculties produce not only adaptive erroneous beliefs, but erroneous beliefs that together form a coherent, uniform experience. Why should we suppose evolutionary processes did all of this extra sophisticated work? Given our ordinary experience, Plantinga's argument requires a type of unreliability that evolutionary theory does not support. Plantinga's account seems to presuppose a beguiling sort of intelligent design in evolutionary processes that any serious naturalist would reject.

The second point concerns Plantinga's claim that the theist is in a better position than the naturalist for explaining the nature and source of our cognitive systems. This claim can be challenged fairly easily through a careful examination of those systems. For example, a number of studies have demonstrated that human cognition is, in certain limited ways, prone to systematic error.[16] In other words, it would appear that a small percentage of our inference strategies actually are unreliable. To use a computer metaphor, our inference programs are not all bug-free.

Now it seems that this observation is somewhat of an embarrassment for the evolutionary reliabilist, who after all claims that natural selection should strongly favor reliable inference strategies. But if these observations are embarrassing for the evolutionary reliabilist, they are also embarrassing for practically *any* version of reliabilism, including theistic varieties which, like

[16] Stephen Stich, "Could Man Be an Irrational Animal?" *Synthese* 64 (1985): 115–135; and *The Fragmentation of Reason* (Cambridge: MIT Press, 1990).

Plantinga's, claim that God designed reliable cognitive mechanisms. Now of these two alternatives—evolutionary reliabilism and theistic reliabilism—we can ask which one does a better job of accounting for these findings.

Well, on the one hand, if naturalism is true, occasional cases of unreliable cognition are to be expected, since on this account cognitive mechanisms come about through unguided evolutionary processes that can sometimes go off in strange directions. While an evolutionary reliabilist will claim that most of our inference strategies are reliable, he can allow for certain pockets of irrationality, produced by variables like the Garcia effect. On the other hand, if you claim that we are made in the image of an infallible and omniscient being, and claim that such a being would never design our cognitive faculties in a way that would lead to systematic error (as Descartes once claimed), then, at least initially, it would seem that these findings are not what one would expect. Hence, it seems that it is really the Christian reliabilist, and not the evolutionary reliabilist, who has some explaining to do regarding the facts of human cognition. As Darwin himself noted, it is the existence of faulty design and imperfections that point to the most plausible origin of biological systems.

Though I have little doubt that a theistic reliabilist can construct some sort of reply to this argument, that discussion will take us too far afield. My main goal has been to convince you that evolutionary reliabilism has a great deal to be said for it, and that it is much harder to undermine that Plantinga's arguments suggest. Neither of Plantinga's two central arguments against evolutionary reliabilism succeed or lead to the conclusion that naturalism is self-defeating. Still, Plantinga's larger argument against naturalism is extremely helpful for those of us who want to defend reliabilism from a naturalistic perspective. Since the status of reliabilism within an evolutionary framework is something naturalists themselves have questioned (see Stich 1990, Stein 1996), we owe Plantinga a debt of gratitude for showing us some of the potential difficulties of giving reliabilism up.

Is Science Biologically Possible?

Jerry Fodor

I hold to a philosophical view that, for want of a better term, I'll call by one that is usually taken to be pejorative: *Scientism*. Scientism claims, on the one hand, that the goals of scientific inquiry include the discovery of objective empirical truths; and, on the other hand, that science has come pretty close to achieving this goal at least from time to time. The molecular theory of gasses is, I suppose, a plausible example of achieving it in physics; so is the cell theory in biology; the theory, in geology, that the earth is very old; and the theory, in astronomy, that the stars are very far away. Scientism is, to borrow a phrase of Hilary Putnam's, the scientist's philosophy of science. It holds that scientists are trying to do pretty much what they say that they are trying to do; and that, with some frequency, they succeed.

I'm inclined to think that Scientism, so construed, is not just true but *obviously and certainly* true; it's something that nobody in the late twentieth century who has a claim to an adequate education and a minimum of common sense should doubt. In fact, however, Scientism is tendentious. It's under attack, on the left, from a spectrum of relativists and pragmatists, and, on the right, from a spectrum of Idealists and a priorists. People who are otherwise hardly on speaking terms—feminists and fundamentalists, for example—are thus often unanimous and vehement in rejecting Scientism. But though the rejection of Scientism makes odd bedfellows, it somehow manages to make them in very substantial numbers. I find it, as I say, hard to understand why that is so, and I suppose the Enlightenment must be turning in its grave. Still, over the years I've sort of gotten used to it.

I'm grateful to Professor Plantinga for some email exchanges about this material, and for allowing me to read his "Naturalism Defeated" in a manuscript version. Plantinga doesn't think I've got to the bottom of the arguments he gives there. No doubt he's right.

Recently, however, the discussion has taken a turn that seems to me bizarre. Scientism is now being attacked by, of all people, *Darwinists*. Not all Darwinists, to be sure; or even, I should think, a near majority. Still, the following rumor is definitely abroad in the philosophical community: When applied to the evolution of cognition, the theory of natural selection somehow entails, or at a minimum strongly suggests, that most of our empirical beliefs aren't true; a fortiori, that most of our empirical scientific theories aren't true either. So the rumor is that Darwinism—which, after all, is widely advertised as a *paradigm* of scientific success (I've heard it said that Darwinian adaptationism is the best idea that anybody's ever had, and that natural selection is the best confirmed theory in science)—Darwinism, of all things, undermines the scientific enterprise. Talk about biting the hand that feeds you!

There isn't, even among those who propagate this rumor, much consensus as to what one ought to do about it; indeed, there's currently a lively philosophical discussion on just that topic. Some say "well, then, so much the worse for truth." Others say "well, then, so much the worse for science." Still others say, "well, then, the paradox must be merely apparent; nothing true can refute itself."

In this discussion, I'll take a different line from any of these. My theme is that we don't need to worry about what to do if Darwinism threatens Scientism because, in fact, it doesn't. There is, I claim, nothing at all in evolutionary theory that entails, or suggests, or even gives one grounds to contemplate, denying the commonsense thesis that scientific inquiry quite generally leads to the discovery of objective empirical truths. I'll try to convince you of this by looking, in some detail, at the arguments that are supposed to show the contrary. I hope you won't find the process tedious; this sort of inquiry can be edifying and, anyhow, it's what philosophers mostly do for a living.

Perhaps I should mention, in passing, that I'm not actually an adaptationist about the mind; I doubt, that is, that Darwinism is true of the phylogeny of cognition. For reasons I've set out elsewhere,[1] I think our kinds of minds are quite likely "hopeful monsters," which is why there are so few of them. It would thus be open to me to take the line that if Scientism is jeopardized by adaptationism about the evolution of the cognitive mind, so much the worse for adaptationism about the evolution of the cognitive mind. But, though I could do that, on balance I'd rather not. That's because I'm much more confident that Scientism is true than I am that Darwinism about the phylogenesis of cognition is false. I want to leave open the possibility that they might *both* turn out to be true. I'm not, therefore, about to concede that they aren't compatible. For present purposes, then, I propose to suppress my qualms and pretend that I believe that our minds evolved; and that they did so under selection pressures. The question of interest is what, if anything, this implies about the likely truth of Scientism.

[1] Jerry Fodor, *In Critical Condition* (Cambridge: MIT Press, 1998), chaps. 14, 15.

So, what is the argument from adaptationism about cognition (henceforth, for short, the argument from "Darwinism") to the rejection of Scientism? Actually, it's an inference that comes in several different packages. I propose to look at two. Here, to begin with, is a recent quotation from Patricia Churchland. "There is a fatal tendency to think of the brain as essentially in the fact-finding business. . . . Looked at from an evolutionary point of view, the principle function of nervous systems is to get the body parts where they should be in order that the organism may survive. . . . Truth, whatever that is, definitely takes the hindmost."[2]

Clearly, the crux of this is the claim that "the principle function of nervous systems is [not the fixation of true beliefs, but] to get the body parts where they should be in order that the organism may survive." Just why does Churchland think that "the evolutionary point of view" implies this? It's not, of course, that there's been some hot new paleobiological discovery she's just heard about ("Flash: Fossil Nervous System Found Functioning in China"). Rather, what she apparently has in mind is some inference from Darwinist general principles. Perhaps it's this: *Organisms* get selected for getting their bodies where they should be; so *nervous systems* get selected for getting the bodies of organisms to where they should be; so *the function* of nervous systems—the function for which selection designed them—*is* to get the bodies of organisms to where they should be. A fortiori, the function of nervous systems isn't "fact-finding."

But if that is the intended inference, it commits a sort of distributive fallacy; it's like arguing that, since the army is numerous, each of the soldiers must be numerous too. Consider the following parody: "Selection pressures favor reproductive success; the design of the heart was shaped by selection pressures; so the function of the heart is to mediate reproductive success." Parallel arguments would show that the function of *every* organ is to mediate reproductive success; hence that every organ has the same function.

Poppycock! The function of the heart—the function that its design reflects—is to circulate the blood. There is no paradox in this, and nothing to affront even the most orthodox Darwinist scruples. Animals that have good hearts are selected for their reproductive success as compared to animals that have less good hearts or no hearts at all. But their reproductive success is produced by a biological division of labor among their organs, and it's the function of the heart *relative to this division of labor* that the heart was designed to perform.

Likewise, it's entirely possible that the kind of mental architecture that maximizes behavioral adaptivity is also one that institutes a division of labor: Perhaps a cognitive mechanism that is specialized to figure out what state the world is in interacts with a conative mechanism that is specialized to figure

[2] Patricia Churchland, "Epistemology in the Age of Neuroscience," *Journal of Philosophy* 84 (1987): 548.

out how to get what one wants from a world that is in that state. That's, roughly, the sort of nervous system that you'd expect if creatures act out of their beliefs, desires, and practical intelligence; and there's nothing in Darwin to suggest that they do otherwise. All Darwin adds is that, if that *is* how a creature works, then, on average, it will work better, and last longer, and have more progeny in proportion as its beliefs are true and its desires are appropriate to its ecological condition. Who would deny it?

Churchland's argument doesn't convince. But maybe there's better on offer? Here's a line that Alvin Plantinga has recently been pushing: "Our having evolved and survived makes it likely that our cognitive faculties are reliable and our beliefs are for the most part true, only if it would be impossible or unlikely that creatures more or less like us should behave in fitness-enhancing ways but nonetheless hold mostly false beliefs."[3] But, in fact, it isn't impossible or unlikely that creatures more or less like us should behave in fitness-enhancing ways but nonetheless hold mostly false beliefs. "[T]he simplest way [of seeing how evolution might get away with selecting for cognitive mechanisms that produce generally false beliefs] is by thinking of systematic ways in which . . . beliefs could be false but still adaptive. . . . [Imagine a creature that] thinks all the plants and animals in [its] vicinity are witches, and [whose] ways of referring to them all involve definite descriptions entailing witchhood. . . . [T]his would be entirely compatible with [its] beliefs' being adaptive; so it is clear . . . that there would be many ways in which [its] beliefs could be for the most part false, but adaptive nonetheless."[4]

It looks like the argument that Plantinga has on offer is something like this:

> *Argument A*: Our minds evolved, so we can assume that our behavior is mostly adaptive. But our behavior could be mostly adaptive even if our beliefs were mostly false. So there's no reason to think that most of our beliefs are true.

At worst, Argument A seems better than Churchland's since it concedes, straight off, that the mind is a belief-making mechanism: Plantinga's point is that you can imagine the mind's succeeding as a belief-making mechanism even if the beliefs that it makes aren't true. Still, I don't think that Argument A could be exactly what Plantinga has in mind; not, at least, if what he has in mind is an argument against Scientism (or against what he calls "naturalism," which, at least for present purposes, we may take to be much the same thing). But I do think that there's quite an interesting argument in the immediate vicinity of Argument A; one that has every right to be taken seriously. We'll

[3] "Naturalism Defeated" (unpublished), 5. There is, in fact, more to Plantinga's argument than what I will discuss here. Plantinga thinks (as I do not) that there is some serious chance it will turn out that we don't act out of the content of our mental states *at all*, a fortiori that the truth of our beliefs can't be what explains our behavioral successes. In any event, I'll consider only the aspects of Plantinga's polemic that turn on Darwinist considerations.

[4] "Naturalism Defeated," 10.

come back and fix this presently. First, however, I want to grumble a bit about a premise.

It's important to Plantinga that it not be in doubt that a system of mostly false beliefs could be adaptive. If there were anything wrong with the notion of a system of mostly false but adaptive beliefs, Plantinga's Darwinian argument against Scientism couldn't get off the ground. Plantinga's way of assuring us that there's nothing wrong with this notion is to sketch a general procedure for constructing such a system: Start with our beliefs, but replace all the referring expressions by complex demonstratives (or definite descriptions) with false presuppositions. We think *that tree is very big*; they think *that witchtree is very big*; we think *the cat on the mat is asleep*, they think *the witchcat on the witchmat is asleep*. And so on. What you end up with is a creature whose behavior is plausibly about as adaptive as ours, but most of whose beliefs are false.

Now I do think that goes a little fast. After all, one needs to explain *why* the behavior of a creature thus epistemically situated would be largely successful; and the intuition, surely, is that, on any way of counting them that's defensible, "most of the beliefs" that mediate its behavioral successes turn out to be true after all. Let it be that *that appletree witch is blooming* is false, or lacks a truth value, in virtue of its presupposing that that appletree is a witch. Still, much of what a creature believes in virtue of which it believes that *that appletree witch is blooming* (and in virtue of which the thought that *that appletree witch is blooming* leads to behavioral successes) are perfectly straightforwardly true. For example: *that's an appletree; that's blooming; that's there; something is blooming; something blooming is there*, and so on indefinitely. The point is trivial enough: If a creature believes *that appletree witch is blooming*, then it presumably believes that *that's an appletree* and that *that's a witch* and that *that's blooming*. And two of these are *true* beliefs that the creature shares with us, and that enter into the explanation of its behavioral successes vis-à-vis blooming appletrees in much the same way that the corresponding beliefs of ours enter into the explanation of our behavioral success vis-à-vis blooming appletrees. Notice that this argument doesn't depend on assuming (as, in fact, I don't) that belief is closed under implication, or even under "immediate" implication. All it needs is the truism that the creature has whatever beliefs the explanations of its behavioral successes require attributing to it. This, surely, is common ground *whether or not* the beliefs that such explanations appeal to are supposed to be true.

The philosopher Berkeley seems to have believed that tables and chairs are logically homogeneous with afterimages. I assume that he was wrong to believe this. But intuition strongly suggests that the truth of "most of" his beliefs survived his metaphysical aberrations, and that it was these residual true beliefs that explain his behavioral successes. He may, for example, have chosen to sit in this chair rather than that one because he believed that this afterimage chair is more comfortable than that afterimage chair. And that may

have been the right choice in the circumstance. If so, we should credit Berkeley's success to his true belief that this chair is comfortable, and not to his false belief that this chair is an afterimage.

The long and short is: There's admittedly no agreed-on way of counting beliefs; but it's not clear that the beliefs of the creature Plantinga asks us to imagine are mostly false. Or, indeed, that its epistemic situation actually differs very much from our own. So, suppose that the forces of selection can't distinguish that creature from us; suppose, that is to say, that its aggregate beliefs are about as adaptive as ours. Even so, so far we don't have a crystal clear case of the kind that Plantinga needs: one where selection might favor a creature whose beliefs are, in some really radical sense, "mostly false."

If I'm right about this, it leaves us at a bit of an impasse. I'm pretty sure that Plantinga's way of constructing a system of beliefs that are clearly mostly adaptive but mostly false doesn't work. On the other hand, I haven't an argument for what I strongly suspect, that there is *no* way of constructing such a system. And I don't want the discussion to break down here, because I think the following question is worth considering: *If it is possible to construct a system of beliefs which would be adaptive though mostly false, is there then an argument from Darwinism to the rejection of Scientism?*

So here's what I propose. Let's just assume that there is a "charmed circle" of nonstandard cognitive mechanisms such that:

- every cognitive mechanism in the charmed circle produces systems of mostly false beliefs, and
- most creatures that act on such systems of false beliefs (most "false believers" as I'll say) will do *at least as well* as most creatures that act on the corresponding true beliefs (i.e., at least as well as most corresponding "true believers").

It is very nice of me to concede all this since I really am much inclined to doubt that it's true.

We need to do one more preliminary thing before we can get to the serious stuff. As we've seen, Plantinga writes that "Our having evolved and survived makes it likely that our cognitive faculties are reliable and our beliefs are for the most part true, only if it would be impossible or unlikely that creatures more or less like us should behave in fitness-enhancing ways but nonetheless hold mostly false beliefs."[5]

Now, that's okay as it's stated because, as it's stated, it only purports to give a *necessary* condition for

(1) Our having evolved makes it likely that our cognitive faculties are reliable.

[5] Ibid., 5.

But what Plantinga actually needs, if he's to have an argument against Scientism, is something like a sufficient condition for

(2) Our having evolved makes it likely that our cognitive faculties are unreliable.

Short of arguing for (2), the most Plantinga's got is that Darwinism is *compatible* with Scientism's being false; which I suppose to be true but not interesting. That the cat is on the mat is *also* compatible with Scientism being false. So what? The cat's being on the mat isn't therefore *a reason to think* that Scientism *is* false. Churchland appears to make a similar mistake: If evolution selected our minds for moving our bodies, then perhaps Darwinism gives us no reason to suppose that our beliefs are mostly true. But that's quite different from Darwinism giving us some reason to suppose that our beliefs are mostly *not* true, and it's this latter that the Darwinian argument against Scientism requires.

In short, what Plantinga really needs (and what I haven't any doubt is what he really intends) is argument A*. Argument A* is just like Argument A except that the conclusion is stronger.

*Argument A**: Our minds evolved, so we can assume that our behavior is mostly adaptive. But our behavior could be mostly adaptive even if most of our beliefs were mostly false. So there is reason to think that most of our beliefs *are* false.

Now, if I were in a bad mood, or if I were generally an unobliging chap, I could, with some justice, harp on the difference between arguments A and A*. Consider:

Argument B: I am standing here giving a lecture and I am wearing my gray socks. But there are other colors (green, say) such that, quite likely, someone a lot like me might be standing here giving a lecture if he were wearing socks of any of those colors. So *that I am wearing gray socks* is not a reason to believe *that I am standing here giving a lecture.*

Argument B seems alright; that I am wearing gray socks *isn't* a reason to believe that I am standing here giving a lecture. It's just that, as a matter of fact, both happen to be true. But, notice, argument B* doesn't seem right at all.

*Argument B**: I am standing here giving a lecture and I am wearing gray socks. But there are other colors of socks (green, say) such that, quite likely, someone a lot like me might be standing here giving a lecture if he were wearing socks of any of those colors. *So that I am wearing gray socks is a reason to believe that I am not standing here giving a lecture.*

Poppycock once again. But, as far as I can see, the difference between B and B* is just like the difference between A and A*, and it's important not to

equivocate between A and A* for just the same reason that it is important not to equivocate between B and B*. I'm a little worried that Plantinga's Darwinian argument against Scientism trades on this sort of equivocation; I haven't, anyhow, found a way to make it run that clearly doesn't.

So, as I say, if I were in a bad mood or generally unobliging I could make a fuss of this. But I'm not going to because, as it turns out, my main line of objection to Plantinga is good against both A* and A. Nothing that follows, therefore, will depend on the difference between them (except at the very end).

Recapitulation: I've agreed to assume that the charmed circle is nonempty. By this assumption, it's likely that there could be false believers that would, in principle, be just as successful as true believers are. What I need to show is that, pace Plantinga, this assumption doesn't even come *close* to underwriting either A or A*.

Well, it doesn't. Here's why: What I've just conceded is, in effect, that what evolution "really wants" is not true believers, but *either* true believers *or* just any old believers whose cognitive mechanisms are in the charmed circle. But the question nonetheless arises, *how, in our case, did evolution go about getting what it wanted?* Darwinism is, remember, a *historical* thesis about *us*. If it's to jeopardize Scientism, it's got to underwrite the claim that our minds actually evolved under selection that favored us *because* our false beliefs were generally successful. That there are, in principle, systems of generally successful false beliefs isn't enough to show this; nor is it enough that there quite possibly could be creatures a lot like us whose minds evolve under selection for false beliefs and whose behaviors are generally successful. What's required, to repeat, is a plausible historical scenario according to which *our* minds were selected because we had lots of behaviorally successful false beliefs. Well, is there such a scenario? The answer depends on the status of the following claim:

> *There is some property P of our beliefs such that:*
> i. In past, our acting on false beliefs that have *P* has actually been (at least) as successful as our acting on true beliefs was. (Otherwise selection would have had a reason for preferring true beliefs to false beliefs with *P*.)
> ii. It's reliable (counterfactual supporting) that false beliefs with *P* are at least as behaviorally successful as true beliefs. (That is, at least some cognitive mechanisms that produce *P* beliefs are in the charmed circle.)
> iii. Assuming our cognitive system is still doing what it evolved to do—that is, still performing the function for which it was selected—our current behavioral successes are *still* largely predicated on false beliefs that have *P*.

The reason there has to be such a property is simply that, according to the present story, evolution selected us *because of* something about lots of our false beliefs that reliably lead to our behavioral success. So the present story

couldn't be true unless there *was* something about lots of our false beliefs that reliably lead to our behavioral successes.

Well, was there? I doubt it, and (short of philosophical skepticism), nobody I've heard of actually believes it. What we all believe is that when actions out of false beliefs are successful, that's generally a *lucky accident;* and, correspondingly, that a policy of acting on false beliefs, even when it works in the short run, generally gets you into trouble sooner or later. Indeed—and this is the point I want to stress—the commonsense view is that we have, even as we stand here now, lots of *solid inductive evidence* that our false beliefs pretty generally don't have property *P*. The inductive evidence is that, though some of our actions on false belief have succeeded from time to time, pretty generally our actions on false beliefs have *failed;* and they've failed precisely *because* they were actions on false beliefs. Conversely, in what must surely be overwhelmingly the standard case, if the explanation of a behavioral success invokes a creature's beliefs and desires at all, it does so as follows: The creature acted in such and such a way because it wanted to bring it about that so and so; and because it believed, truly, that acting that way *would* bring it about that so and so. That much about intentional explanation has been clear since Aristotle. Perhaps, when you're doing philosophy, you're prepared to doubt it; but you don't, surely, doubt it when you're being serious.

This commonsense view of the relation between true belief and successful behavior may be wrong, of course, but I want to stress that you can't invoke the *assumption* that it's wrong on behalf of the putative Darwinian argument against Scientism. For, that we are probably false believers is supposed to be the *conclusion* of the Darwinian argument, not its *premise*. It is very easy when one is doing epistemology—especially when one is doing skeptical epistemology—to find oneself arguing backwards. I want to pause to rub this point in a bit, because I think it may really be at the bottom of a lot of the trouble.

You clearly could get the required antiscientistic Darwinian conclusion from the following argument *assuming that it's premises are true.*

> *Argument C:* 1. Many (most) of our behavioral successes probably are (/probably have been) predicated on false beliefs that have *P*. 2. So selection has had grounds for favoring false beliefs that have *P*. 3. So we've probably evolved to have false beliefs that have *P*.

But, surely, Argument C is question begging in the present context. As it stands it's no better than the first premise, and on the face of it, the first premise is not plausible. What's needed, to make the argument convincing, is some *independent* reason for thinking that many of our behavioral successes actually have been predicated on false beliefs that have *P*. And, on pain of circularity, it better be a reason that doesn't itself assume that we've probably evolved to have false beliefs. Short of philosophical skepticism, I know of no independent reason to believe premise 1 and, as remarked above, all the inductive evidence I've heard of is to the contrary.

I hope the main line of my argument against Plantinga is now getting clear. It goes in three steps.

Step 1: Believing that there is no such property as *P* is *not at all* the same as believing that the charmed circle is empty, or believing that it's *likely* that the charmed circle is empty. I, for example, believe that, in point of historical fact, those of my successful actions that have been predicated on false beliefs were mostly flukes; a fortiori, that they weren't predicated on false beliefs that had *P*. But I also believe that, very likely, there could be cognitive systems of the sort that membership in the charmed circle requires. All *that* needs is some property such that *if* my false beliefs had had it, my actions *would have been* reliably successful. It's quite consistent of me to believe that there very likely *are* properties such that if my false beliefs had had them they would have been reliably successful, but to doubt that the false beliefs on which my successful actions have *actually* been predicated actually *have* had any of these properties. And likewise for the false beliefs on which the successful actions of my grandmother were predicated; and likewise for those of her grandmother; and likewise back to the protoplasmic slime.

Step 2: For it to be true that I was selected because in the past my false beliefs had *P*, it would at least have to be true that lots of my past behavioral success have been predicated on false beliefs tout court. But, in fact, I have no reason at all to believe that lots of my past behavioral successes have been predicated on false beliefs, and I have vast lots of inductive evidence to the contrary. For example, this morning I managed to get my teeth brushed; a small behavioral success, to be sure, but mine own. Such as it was, I'm quite certain that it was prompted and guided by a host of true beliefs including, inter alia: true beliefs about my teeth needing a brush; true beliefs about the spatiotemporal location of my toothbrush, true beliefs about the spatiotemporal location of my teeth, true beliefs about the spatiotemporal location of my limbs with respect to my toothbrush and my teeth, and so on. Certainly, short of philosophical skepticism, I can think of no reason in the world to deny any of this.

I pause to emphasize the caveat. I said that I can think of no reason *short of philosophical skepticism* for doubting the commonsense view that most of one's successful actions are and have been actions out of true beliefs. This way of putting it is *not* question begging in the present context. For, I am not trying to refute philosophical skepticism, nor do I doubt that Darwinism *together with* philosophical skepticism would argue that we probably aren't true believers. That must be so since philosophical skepticism, all by itself, would argue that we probably aren't true believers if only *it* were true. None of that is pertinent, however, in the present context; the question before the house is whether Darwinism undermines Scientism from, as it were, the inside; namely, *without* supplementary philosophical commitments.

Step 3: If the mere nonemptiness of the charmed circle proves nothing, and if the available inductive evidence is that our successful actions are quite generally predicated on true beliefs, then all that's left for the Darwinian argu-

ment against Scientism is to suppose that things have changed a lot since our minds got selected: Whatever the situation is *now*, it *used to be* that our successful behavior was generally predicated (not on true beliefs but) on false beliefs that have *P*. That's how it was back when selection molded cognition.

But why on earth should anyone believe this? And, even if anyone does believe it, why does it matter in the present context? That minds used to achieve their successes by causing us to act on false beliefs is no challenge to Scientism *as long as they don't do it that way any more*. The relevant consideration is one that Gould and Lewontin stressed in their famous paper about spandrels: "[there are cases where one finds] adaptation and selection, but the adaptation is a secondary utilization of the parts present for reasons of architecture, development, or history. . . . If blushing turns out to be an adaptation affected by sexual selection in humans, it will not help us understand why blood is red."[6] Correspondingly, even if it turns out that nervous systems were originally selected for making adaptive false beliefs—a hypothesis which I don't believe for a minute, and which we've thus far been offered no reason to grant—that *still* wouldn't ground a Darwinian argument against Scientism.

Now we're ready to recap. The polemical situation, according to the opposition, is that a Darwinist has no reason for supposing that evolution picks true believers rather than false-believers-with-cognitive-mechanisms-in-the-charmed-circle. This is because, by assumption, the behavior of the latter would be just as adaptive as the behavior of the former. But, I claim, that's not good enough. It can't be plausible that we were selected for the behavioral success of our false beliefs unless it is *independently* plausible (or, anyhow, independently not *im*plausible) that a lot of our behavioral successes have in fact been actions out of false beliefs. But, to repeat, the commonsense view is that, however we may have evolved, our behavioral successes are generally predicated on *true* beliefs; and that it's generally the truth of the beliefs on which they are predicated that explains the successes of the behaviors; and that, in cases when we have managed to act successfully out of false beliefs, that's mostly been blind luck. As far as I can tell, there is no reason on God's green earth why a proper Darwinism should reject this commonsense view.

I propose to conclude with a caveat. It's been one of my complaints against Churchland and Plantinga that they seem to want to argue from "Darwinism doesn't imply that most of our beliefs are true," to "it's therefore likely that most of our beliefs are false"; an argument which is, as I remarked, invalid on the face of it. But I also want to warn against the converse fallacy, that of arguing from "Darwinism doesn't imply that most of our beliefs are false" to "it's therefore likely that most of our beliefs are true."

A fair lot of philosophers have been tempted by the idea that, far from re-

[6] S. J. Gould and R. C. Lewontin, "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Program," Proceedings of the Royal Society of London, ser. B, Biological Sciences, vol. 205, no. 1161, *The Evolution of Adaptation by Natural Selection* (21 September 1979): 593.

futing Scientism, the theory of evolution actually underwrites it. Thus Dan Dennett: "[T]he capacity to believe would have no survival value unless it were a capacity to believe truths."[7] What he's endorsing is not the casual truism that any creature that is able to believe falsely that *P* is likewise able to believe truly that not-*P*; it's the much stronger claim that the Darwinian survival of a thinking creature ipso facto implies, or anyhow makes it plausible, that most of what that creature believes is true.

But that simply isn't so, and the reason is much the one that Plantinga suggests. That a creature evolved to have some organ is, if the Darwinian story is true, something like a sufficient argument that having that organ was adaptive for the creature. But it is not part of the Darwinian story, nor is it independently the case, that if a creature *failed* to evolve to have some organ, that's anything like a sufficient argument that having it *wouldn't have been* adaptive for the creature. Darwinism deals in *sufficient* conditions for adaptivity, not in *necessary* conditions for adaptivity.[8] The point this observation turns on is, in fact, perfectly general: Darwinism isn't about which traits of organisms *would have been adaptive if a creature had had them*. Darwinism is about what effects the actual possession of a trait has as a function of its adaptivity: roughly, the frequency with which a trait is represented in a population is proportional to its adaptivity, all else equal.

To put much the same point the other way round: What Darwinism requires as (roughly) sufficient for the evolution of a trait is *both* that the trait would be adaptive if the organism were to have it; *and* that, in point of historical fact, there are some (rudimentary, approximate, proto-) instances of the trait actually on the ground for selection to operate upon. In particular, it is quite compatible with Darwinism that our cognitive mechanisms might have been selected to produce adaptive false beliefs, assuming there were lots of those on the ground when our minds evolved. I've supposed, however, that what was on the ground when our minds evolved were primitive cognitive mechanisms that produced, by and large, adaptive true beliefs; and that selection favored, among such cognitive mechanisms, the ones that produced adaptive true beliefs efficiently and reliably and about a wide range of ecologically salient states of affairs. The reason I've supposed this is that, pretty clearly, it's mostly true beliefs that eventuate in adaptive behavior *now*.

But, to repeat, Darwin would have been just as happy if what had been on offer for selection to choose from when we evolved was cognitive mechanisms which produce, by and large, adaptive *false* beliefs; in that case, the Darwinian prediction would have been that it's mostly false beliefs that eventuate in adaptive behavior now. It is, just as Churchland and Plantinga both

[7] Daniel Dennett, "Intentional Systems," in *Brainstorms* (Cambridge: Bradford Books/MIT Press, 1978), 17.

[8] This is quite a general property of teleological explanations; as, by the way, Carl Hempel pointed out some while ago (C. G. Hempel, "The Logic of Functional Analysis," in *Aspects of Scientific Explanation* [New York: Macmillan, 1965]).

say, the adaptivity of the beliefs, and not their truth per se, that evolution cares about. Since, for all we know, evolution would have chosen false believers if it had been given the chance, the fact that evolution chose *us* isn't, in and of itself, a reason for thinking that we're true believers. There is, as far as I can tell, *no* Darwinian reason for thinking that we're true believers. Or that we aren't.

The problems of philosophy are very hard, so we keep hoping that somebody else will take them off our hands. Darwin, Heaven help him, has been for years among our favorite candidates. In the nineteenth century, we thought that maybe he would do our ethics for us. These days, it's the burden of grounding our epistemology that we want him to carry. According to temperament, we'd either like him to show that most of our beliefs have to be true or that most of them have not to be. Either way, we want Darwin to demarcate, once and for all, the necessary limits of human fallibility. He's done so much for biologists; why can't he do this one little thing for philosophers?

But Darwin isn't in the epistemology business, and evolution *doesn't care* whether most of our beliefs are true. Evolution is *neutral* as to whether most of our beliefs are true. Like Rhett Butler in the movie, it just doesn't give a damn.

So, then, why do I believe in Scientism? Or, to put the question slightly differently: If the theory of natural selection doesn't imply that our empirical beliefs are generally pretty close to being true, then why, in particular, do I believe that the lunar theory of the tides, and the tectonic theory of the arrangement of the continents, are pretty close to being true? Put that way, the question sort of answers itself. I believe these theories on account of my (casual, to be sure) acquaintance with the evidence that has been alleged for them. And, of course, the evidence that's been alleged for the lunar theory of the tides isn't that human cognition evolved; it's that, for example, the lunar theory correctly predicts that (all else equal) the tides are highest when the moon is full. Likewise, the evidence alleged for the tectonic theory of the arrangement of the continents isn't that human cognition evolved; it's that, for example, the geology of Western Africa matches the geology of Eastern South America in ways that are otherwise quite inexplicable. This is just what you'd expect; and it's just what the scientists themselves take for granted. What on earth does whether or not our minds evolved have to do with whether or not it's the moon that pulls the tides around?

Maybe, though I very much doubt it, the whole scientific enterprise somehow requires, or presupposes, an epistemological foundation to rest on. And if philosophers can't get it one a priori, by an exercise of Cartesian intuition, maybe they can get it one some other way. Well, maybe. But Darwin isn't going to help. On the contrary, since the theory of evolution is itself a part of the scientific enterprise, whatever epistemological boat we're in, Darwin's in there with us. My Scientism is not in the least disturbed by this. Quite the opposite, it's perfectly delighted to have him aboard.

Darwin's Doubt, Calvin's Calvary

Evan Fales

In the closing chapter of his book *Warrant and Proper Function*, Alvin Plantinga turns the tables on the theories of knowledge generally described as naturalized epistemologies. What is ingenious about the attack is that Plantinga's own epistemology is fairly described as a "naturalized" one. Like its rivals, it invokes conditions external to the knower which bear upon the truth-conduciveness of her cognitive mechanisms. Nevertheless, there is a profound difference between Plantinga's naturalized epistemology and its rivals. One motivation for naturalism, shared by Plantinga, is the need for a Gettier-proof theory.[1] But a second aim which has driven naturalism is not one for which Plantinga has any sympathy. Since its beginnings, naturalistic epistemologists have wanted a theory of knowledge which can be fitted within a scientific, preferably materialist, worldview—and correlatively, within a naturalistic ontology. But with an ironic twist, Plantinga argues that a naturalized epistemology can only succeed when hitched to a supernaturalistic—specifically, a theist—ontology. Thus, Plantinga's target is not naturalized epistemology *simpliciter*, but rather naturalized epistemology coupled with ontological naturalism. I shall take naturalism[2] to be the thesis that there exists nothing other than spatiotemporal beings embedded within a space-time framework. Thus, naturalism does not countenance, inter alia, disembodied minds, gods, and the like.

This essay is a substantial revision of "Plantinga's Case against Naturalistic Epistemology," *Philosophy of Science* 63 (September 1996): 432–451.

[1] Whether Plantinga's theory in fact succeeds in escaping Gettier counterexamples has generated a considerable literature. See, e.g., Fales, "Review of Alvin Plantinga, *Warrant and Proper Function*," *Mind* 103 (1994): 391–395, and various papers in *Warrant in Contemporary Epistemology*, ed. Jonathan Kvanvig (Lanham, Md.: Rowman and Littlefield, 1996).

[2] Plantinga calls this view metaphysical naturalism.

Plantinga's strategy is to argue that naturalism requires the neo-Darwinian theory of evolution (or something closely similar) to account for our existence, but that this theory undermines a naturalistic epistemology in this way: if the neo-Darwinian account is true, then it is unlikely or at least not demonstrably likely that our cognitive faculties are reliable generators of true beliefs. Hence, if naturalism-plus-naturalized epistemology (hereafter NNE) is correct, then we could not be in a position to know that it is, and so could not know that our sensory faculties and inferential processes yield knowledge. In that case, NNE is an unstable position: it gives way to skepticism.

On the other hand, says Plantinga, theism plus naturalized epistemology is a stable position that can resist skepticism. For if God created us, either directly or indirectly through directed evolutionary means, then we have a guarantee that our cognitive faculties, when normal and properly used, are reliable.

Plantinga's argument, then, hinges on three crucial claims: that naturalism is committed to a neo-Darwinian theory of our origins; that neo-Darwinian evolutionary processes would, for all we know, most likely yield cognitive processes which are unreliable in the relevant epistemic sense; and that theistic creation, in contrast, would yield reliable cognitive mechanisms.

It is just here that many supporters of NNE have seen in the neo-Darwinian theory an important ally. They have assumed that natural selection would tend to select for reliable cognitive mechanisms. Natural as this assumption appears, rejecting it lies at the heart of Plantinga's attack on NNE. If Plantinga can make good his claim that NNE is (probabilistically) *in*coherent, whereas theistic naturalized epistemology (hereafter TNE) is coherent, he can claim a significant advantage for TNE. I shall show that this challenge fails. I shall take up each of Plantinga's central theses in turn, but my discussion of the first and third will be rather brief.

I. Are Naturalists Committed to Darwinism?

Plantinga's attack upon neo-Darwinism isn't confined to *Warrant and Proper Function*. In a related paper, "When Faith and Reason Clash: Evolution and the Bible,"[3] Plantinga outlines what he takes to be both empirical and specifically Christian reasons to reject neo-Darwinism.[4] Both there and in *Warrant and Proper Function*, Plantinga repeatedly emphasizes the claim that,

[3] This is the lead paper of a symposium published in *Christian Scholars Review* 21 (September 1991).

[4] To simplify, I shall use the term 'neo-Darwinism' to refer to any theory of biological evolution that has strong affinities to Darwin's and incorporates molecular genetics: specifically, that does not invoke any supernatural agent or purposive cause directing the process, and that takes random variation, the genetic transmission of traits, and natural selection to be central mechanisms.

for naturalists, neo-Darwinism is "the only game in town." He emphasizes this because he wants to show that naturalism falls if neo-Darwinism does. But although he is in a sense quite right about this, his way of putting the point is tendentious, in two respects.

First, it is certainly not true that the falsity of neo-Darwinism *entails* the falsity of naturalism. There is no a priori connection between the two. So far as the a priori possibilities go, anything might be the cause of life as we observe it. However, *relative to what we (think we) know about nonbiological processes in the universe*, it is plausible to claim that neo-Darwinism holds out the only reasonable hope for a nonpurposive account of the existence of life. If we allow also that a purposive account would have to have recourse to some supernatural agent, then Plantinga has his conclusion.

At the same time, it is striking that Plantinga says almost nothing about the *converse* thesis, viz., that Christian theism (at least Plantinga's favored version of it) becomes implausible if neo-Darwinianism is *true*. He emphasizes the stake that naturalists have in the truth-value of neo-Darwinism, while remaining virtually silent about the risks for TNE. This is the second (and more serious) way in which Plantinga's point is tendentiously put.

II. Is the Reliability of Cognitive Faculties Improbable, Given Neo-Darwinism?

My main aim in this chapter is to argue that Plantinga has not demonstrated that this is so. If Plantinga's argument fails here, then he will not have shown that NNE is probabilistically incoherent. Therefore he will not have shown that a consistent naturalistic epistemology must be framed within a supernaturalistic ontology.

Plantinga proceeds as follows. First he softens the enemy position by citing the work of philosophers—most notably Stephen Stich[5]—who have argued, on various grounds, that evolutionary processes should not be expected regularly to lead to reliable cognitive mechanisms (and in fact often do not). (Stich, in turn, relies heavily upon the findings of Kahneman, Slovic, and Tversky[6] and of Nisbett and Ross,[7] which show—quite embarrassingly—how robustly certain fallacious forms of inference are entrenched within the human information-processing system. I shall have more to say about these findings presently.)

A central point for Plantinga—as for Stich—is that natural selection does not directly favor true belief. Rather, it favors appropriate action. Since true

[5] Stephen Stich, *The Fragmentation of Reason* (Cambridge: MIT Press, 1990).

[6] Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press, 1982).

[7] Richard E. Nisbett and Lee Ross, *Human Inference: Strategies and Shortcomings of Social Judgment* (Englewood Cliffs, N.J.: Prentice-Hall, 1980).

beliefs need not engender successful action, nor false beliefs be fatal, naturalists need to show that, on average, reliable belief-forming mechanisms confer an advantage over various alternative possibilities. Indeed, Plantinga is able to haul out a passage in a letter of Darwin's to W. Graham that supposedly reflects this worry: "With me the horrid doubt arises whether the convictions of man's mind, which have been developed from the mind of the lower animals, are of any value or at all trustworthy. Would anyone trust in the convictions of a monkey's mind, if indeed there are any convictions in such a mind?"[8] The citation allows Plantinga to dub this source of the problem for NNE 'Darwin's Doubt'.[9]

In the light of this general reflection, Plantinga proceeds as follows. First he outlines five different evolutionary scenarios, consistent with neo-Darwinism, which would result in our cognitive faculties being unreliable, and which are compatible with most of our beliefs being false. Then, in a preliminary argument, Plantinga suggests that, relative to neo-Darwinism, the objective probability of some one of these scenarios having been realized in our own case is larger than that of a scenario under which evolution has conferred on us reliable belief-generating mechanisms. Therefore, we ought to believe, if we are neo-Darwinians, that our cognitive mechanisms are probably unreliable.

A second argument—Plantinga calls it his main argument—concedes that estimating the objective probabilities here is a risky business. Perhaps the probabilities are not as Plantinga supposes: perhaps, objectively speaking, the mechanisms of neo-Darwinian evolution do tend to favor the development of reliable cognitive mechanisms, when they favor the evolution of cognition at all. Admitting that, Plantinga reasons to a weaker conclusion: in that case, we at least are not in a position to *know* that the probability of reliable cognitive faculties is high, given neo-Darwinism. Weaker, but strong enough. If the reliability of our cognitive powers remains seriously in doubt, given neo-Darwinism, then we cannot appeal to that theory to support a naturalistic epistemology that certifies our most cherished beliefs. In particular, our ground for confidence in neo-Darwinism is itself undermined, and the position becomes probabilistically incoherent or unstable.

Now, in response to this, a naturalist has available two strategies. He or she can, first, simply refuse to address the second-order question, the question whether our cognitive faculties indeed *are* reliable. Naturalists can insist that

[8] Charles Darwin, *The Life and Letters of Charles Darwin* (New York: D. Appleton, 1919), 1:281. But this involves a serious and tendentious misreading of Darwin's intent. For details, see Fales, "Plantinga's Case against Naturalistic Epistemology," *Philosophy of Science* 63 (1996): 436 n. 6.

[9] See Plantinga, *Warrant and Proper Function* (New York: Oxford University Press, 1993), 219–228. As we shall see, there are in fact good empirical reasons to trust, in general, the beliefs of monkeys.

they don't know whether those faculties are reliable or not; nor need this be known in order for it to be the case that many (or most) first-order beliefs are known to be true. That, after all, is just what externalism permits.

However, if a naturalist accepts neo-Darwinism (as is likely), then this response will no doubt not be very satisfying. For one thing, externalists typically *do* want to go on to certify the reliability of our cognitive faculties. For another, even though this response addresses Plantinga's "main" argument, it will not do as a reply to his preliminary argument. If Plantinga has succeeded in *showing* that, given neo-Darwinism, it's unlikely that our cognitive faculties are reliable, then a neo-Darwinian externalist epistemology is indeed in trouble, not with respect to the analytic task of saying what it *is* to know or be justified, but with respect to providing a framework in which our beliefs—from common sense to science—can coherently be affirmed to *be* justified.

The naturalist's best strategy, therefore, is to show that Plantinga is *wrong* in his estimate of the implications of neo-Darwinism for our evaluation of our cognitive powers. And that is what I shall now argue.

There are, Plantinga suggests, five different ways in which neo-Darwinian evolution can produce creatures with belief-generating mechanisms that would offer no comfort to reliabilist conceptions of knowledge.

> (1) The first possibility is that there might be no causal connection between beliefs and action at all. Since the only thing survival and procreation demand is getting one's body parts into the right places at the right times, we can imagine creatures whose adaptive responses to their environment are handled in an entirely cognition-free way, while their beliefs are on permanent holiday. In such creatures, the unreliability of belief-forming mechanisms would be no liability at all (nor would reliability confer a selective advantage).
> (2) A second, related possibility is that beliefs are causally connected with behavior, but only by way of being effects of that behavior, or "side-effects" of the causes of behavior. Here again, the truth or falsity of a belief will play no role in the appropriateness of the behavior it is linked to; and so truth will confer no selective advantage.
> (3) Beliefs might indeed causally affect behavior, but do so in a way that is sensitive only to their syntax, not to their content or semantics. Then once again, the *truth-value* of a belief would be irrelevant to its role in producing adaptive behavior.
> (4) Perhaps beliefs could be causally efficacious, and their content materially relevant to the behavior they help generate, while the behavior thus generated is *maladaptive*. No organism, perhaps, is *perfectly* efficient, perfectly attuned to its environment. Maladaptive characteristics are a burden which can be borne provided they are not *too* maladaptive. Natural selection can even *favor* maladaptive traits. This can happen when two genes, one harmful but the other beneficial, lie closely adjacent on a chromosome, so that they tend to travel together during genetic reshuffling; or it can happen by way of pleiotropy—a single gene coding for two or more traits. Indeed, adjacency and pleiotropy can be invoked to pro-

vide mechanisms that would explain possibilities 1 and 2 as well. The present point is that an organism may be able to hobble along with lots of false beliefs and misdirected actions.

(5) Finally, evolution might produce organisms in which false belief leads to *adaptive* action. As Plantinga points out, this can happen in several ways. Freddy the caveman may believe that saber-toothed tigers make great pets, may want to tame the one that has just appeared, and believe that running away from it as fast as he can is the best way to corral it. Or Freddy may want to be eaten, believe correctly that this cat will eat him if given the chance, and believe falsely that shoving a firebrand in its face maximizes the chances for his desired fate. The general point is that when beliefs cause action, the action that results is the combined result of various desires and beliefs. Thus, a false belief can "cancel out" another false belief, or "cancel out" a destructive desire, to produce a beneficial action. Why should we suppose, then, that if nature selects for creatures whose actions are guided by their beliefs (and desires), it will be *true* beliefs that confer the greatest selective advantage?

Here we have five ways in which the random processes of neo-Darwinian evolution might produce creatures with belief-generating mechanisms whose reliability one wouldn't want to bet on. Perhaps we're such creatures. The other possibility is the evolution of creatures whose actions are guided by constructive desires informed by largely true beliefs and sound inferences. But so far, the chances of this happening are perhaps only one to five—unless the odds favoring possibilities 1–5 are small.

Let's turn, then, to an evaluation of those odds. Since rather similar considerations apply to (1), (2), and (4), I shall take up these possibilities first, then turn to (3) and (5).

The neural systems by means of which organisms generate and manage their beliefs are biologically expensive.[10] Both in terms of the genetic coding required, and in terms of energy expenditure devoted to growth and maintenance, neural systems—brains in particular—are costly devices.[11] It is hard to see how it could be otherwise, given the nature and complexity of their functions. This appears to be especially true of the mechanisms for belief formation and processing. First, beliefs must be generated—in our case, by distinct mechanisms linked to the dozen or more sensory, proprioceptive, and introspective modalities with which we are equipped. Next, to be of use, they must

[10] It is controversial how many animals *have* beliefs, and how sophisticated the reasoning based on those beliefs is. My own view is that many mammals and birds have action-guiding beliefs (or something rather like beliefs) and inferential capacities of varying sophistication. Although it is, notoriously, difficult to formulate the criteria we use to ascribe beliefs to other minds, it is certain that our normal belief-ascribing practices extend naturally, albeit with some attenuation, to many nonhuman species. For confirming data, see, e.g., Frans de Waal, *Good Natured: The Origins of Right and Wrong in Humans and Other Animals* (Cambridge: Harvard University Press, 1996). To avoid undue controversy, I shall confine my discussion to the primates.

[11] See, e.g., Katherine Milton, "Diet and Primate Evolution," *Scientific American* 269 (1993): 90. Milton details the costs—and the advantages for foraging strategies—of larger brain size in spider monkeys.

be catalogued and stored for efficient retrieval, they must constantly be squared with one another to insure as much consistency and inductive coherence as possible, and there must be inferential mechanisms devoted to the production of further beliefs.

Of course, many organisms get along without all of this—or any of it; but *having* these capacities, as we clearly do, is an expensive proposition, biologically speaking. That gives the neo-Darwinian a prima facie reason for assigning a low probability to the development of such mechanisms unless they confer a decided selective advantage. The selective advantage of intelligence, when linked in an appropriate way to action, can hardly be denied. While many ecological niches can be successfully filled without the benefit of intelligence (witness the cockroach), Plantinga will agree that *Homo sapiens* has, more than any other species, specialized in intelligence as a survival strategy. We have few other biological advantages; most of our eggs are in that basket.[12] Our heavy investment in big brains and otherwise mediocre bodies makes it all the more unlikely that resources would be wasted on elaborate belief-forming and processing mechanisms that have no practical utility.

Let's remind ourselves that the question before us is this: given neo-Darwinism, and given the cognitive mechanisms found in *Homo sapiens* and their ecological context, what is the probability that those mechanisms are either maladaptive—(4)—or adaptively irrelevant—(1) or (2)? Very small, say I. But perhaps, as Plantinga suggests, there is a significant probability that Darwinian evolution would produce one of these possibilities via the mechanisms of association or pleiotropy. Yet even granting that association and pleiotropy are common, the occurrence of costly traits that ride piggy-back upon beneficial ones is quite rare, thanks to selective pressures.

Even worse, human intelligence is not a single trait—that is, it is not encoded by a single gene, but by many, many genes. What is the likelihood that these genes—each finely tuned to work in concert with the others to produce an integrated cognitive apparatus—would *all* be getting free rides from other genes and thus be relatively insulated from adaptive pressures?

In fact, there is *no* plausible neo-Darwinian scenario to back up (1), (2), or (4): they are wildly improbable. For the development of something as complex as the human brain, only the slow, tortuous route of selection in virtue of positive step-by-step adaptive change is at all plausible.

What, then, of possibility 3? It allows that beliefs play a role in action but lets this role be quite independent of their content and truth-value. The issues raised by this possibility are more complex. But a central question is

[12] It is generally agreed that the advantages conferred by our erect posture and manual dexterity are closely linked to our intellectual capacities. Our investment in intelligence as a species is extensive. It includes two related and biologically very costly adaptations, namely, giving birth at an earlier stage of fetal development than is the case for any other mammal (so the large skull can pass through the birth canal), and a greater expenditure of energy in childrearing than characterizes any other mammal.

whether it is conceptually possible that semantic content would fail to map onto syntax so as to preserve truth. That is, we must ask: In virtue of what is it the case that a given mental representation denotes or picks out high temperature (say) as opposed to moderate temperature? Here much will depend upon one's views about intentionality. But an entirely reasonable view, from a naturalistic perspective, is that mental representations get their content in virtue of being caused in the right way by items in the environment; and that this is a *conceptual* truth. Thus if a mental representation is caused in the right way by heat, then it's a representation of heat; and if it is not so caused, then it's not a representation of heat. So long as representations are causally linked to the world via the syntactic structures in the brain to which they correspond, this will guarantee that syntax maps onto semantics in a generally truth-preserving way. The details here are unavoidably complex; they have to do with how reference to particulars and to properties initially gets fixed, and with what it is for a representation to be caused "in the right way."[13] But from a naturalistic perspective, the probability of (3), while harder to assess, may well be no higher than that of (1), (2), or (4).

The most interesting case, in a way, is (5). According to (5), beliefs are (partly) responsible for action, but in a way that permits actions to be on target while the guiding beliefs are wildly wrong. How plausible is this?

It appears that Plantinga has in mind here a conception of the connection between beliefs, desires, and action that is roughly in keeping with folk psychology. Beliefs and desires function as premises in a bit of practical reasoning; the outcome is a decision to act (or the action itself). Keeping the inference-performing part of our rational processes intact, we are to imagine the production of felicitous action from corrupt beliefs. Of course, there's no reason to assume the inferential mechanisms are sound: maybe they are corrupt as well. Either way, do we have here a workable action-guiding system?

I think not. In the first place, Plantinga's examples, like my Freddy cases, work only because the beliefs which supply the immediate doxastic input to Freddy's practical syllogism are perceptual—that is, noninferential—beliefs, or simple inductive generalizations. That makes it easy to imagine a cognitive mechanism that takes input information, systematically reverses truth-value, and thereby produces *systematically* false beliefs. But what happens when de-

[13] If the semantic content of beliefs were not hooked in some way to the syntactic structure of neural states, then (causally speaking) beliefs would float free, and possibility 3 would degenerate into 1 or 2. For further details on how content is conceptually linked to the causal story, see Fales, *Causation and Universals* (New York: Routledge, 1990), chap. 12. For more on the grounds for rejecting syntactic theories, see Alvin Goldman, *Epistemology and Cognition* (Cambridge: Harvard University Press, 1986), 170–173. Motivation for syntactic theories derives, in part, from externalist, referential semantics and examples such as Hilary Putnam's Twin-Earth cases ("The Meaning of Meaning," in *Language, Mind and Knowledge*, ed. Keith Gunderson, Minnesota Studies in Philosophy of Science, no. 7 [Minneapolis: University of Minnesota Press, 1975]). But such considerations hardly show that belief formation does or could ignore truth-conditions; and indeed, as I've argued, this style of semantics helps to *insure* that cognitive mechanisms will be truth-conducive.

ductive inference comes into play? *True* premises guarantee true conclusions: so a system that relies consistently upon true inputs to guide inference and action can employ general rules and hope to get things (i.e., action) right. But when a deductive argument employs false premises, the truth-value of the conclusion is *random*.[14] Thus there *cannot* be any set of *general* algorithms which get a creature to use the conclusions of such arguments in a way that reliably promotes successful action. A cognitive system which is not *extremely* limited in the inferential procedures it employs must either give up all hope of successfully directing action or become unintelligibly complex and ad hoc in its procedures for connecting belief to action.

But this conclusion applies *even when the beliefs in play* are noninferential or based only on enumerative induction. Freddy, who is carrying a heavy rock he falsely believes to be light and soft, nearly steps on a puff adder. Believing that being hit by something light and soft will be fatal for the adder (also false), he quickly drops the rock on it, and lives to see another day. So far so good for Freddy. Continuing on with his rock, Freddy encounters an angry warthog on the trail. Still believing the rock to be light as a feather, and believing (falsely) that dancing upon something light deters warthogs, Freddy proceeds to do a two-step on top of the rock directly in the path of the charging pig. The moral of this fable is plain: there are no effective algorithms connecting false belief to appropriate action, as there are when the input is true beliefs and the rules of inference employed are valid or inductively sound. Intelligent action is hard enough for a brain to manage; burdening it with ever-changing, completely arbitrary principles would make the task impossible. Freddy may survive the adder, but he'll not live long. Nor will his genetic heritage.

The large and complex primate brain that is required to process information efficiently and flexibly is an extremely expensive adaptation, biologically. There is abundant evidence that, in order for primates to exploit their ecological niche effectively, they must employ cognitive mechanisms that are as cost effective as possible. The array of special mechanisms and algorithms required to satisfy Plantinga's case 5 could not possibly do this.[15]

To Plantinga or anyone else who may think that primates could have achieved these adaptations via belief-generating mechanisms that yield a large number of false beliefs and memories, I offer the following challenge. Construct *in detail* an account according to which a monkey, presented, say, with a fruit-bearing tree, forms (obeying some information-processing algorithm, we must assume) a false belief (*which* false belief?) such that it meets

[14] Arthur Falk has pointed out to me that the randomness can be avoided by employing inference rules that are duals of the valid ones: they take falsehoods into falsehoods. But this still avails naught when it comes to moving from false premises *and benighted desires* to life-saving praxis. Moreover, whereas the *true* beliefs relevant to a given practical deliberation are often obvious (e.g., 'That is a warthog'), which *false* beliefs could effectively serve a dual inference presents a typically hopeless problem ('That is *not* a warthog' is far less informative than its denial).

[15] For more details, see Fales, "Plantinga's Case against Naturalistic Epistemology," 443–444.

the following conditions: (1) combined with other present false beliefs and/or destructive desires, it leads (with good probability) to felicitous action; (2) when combined on other occasions with yet *other* false beliefs/bad desires, it *still* is likely to produce correct action; and (3) if destructive desires are invoked, a plausible Darwinian story can be told about how they evolved from the action-guiding desires of the pre-rational ancestors of the monkey, or in some other way. Now construct a system of algorithms that will achieve this for the monkey's beliefs generally. I say it can't be done.[16]

But perhaps I have overstated the case. Surely—as Nisbett, Kahneman, and others have amply demonstrated—it isn't true that human sensory faculties and inferential processes are in all circumstances reliable. The literature on the fallibility of sense perception is vast and fairly familiar; here, let me focus on inference. What the experiments of Nisbett and others have shown is that certain sorts of inferences, though fallacious, are regularly made. The failings are quite robust: even highly intelligent people fall prey to them (which gives Nisbett and Ross occasion to wink at their fellow academics). Of course, it's not the case that *everyone* succumbs to the fallacies *all* the time; if *that* were so, we'd never know the reasoning to be fallacious. But nevertheless, Stich's question remains: if neo-Darwinian processes of evolution are so effective at winnowing epistemic wheat from chaff, why are these fallacious inferences apparently "hard-wired" into our cognitive makeup? I shall suggest that there are plausible Darwinian strategies for answering that question. But before coming to that, let's have a look at Plantinga's own supernaturalistic version of naturalistic epistemology. It's time to turn the tables on TNE.

III. Is Human Cognitive Reliability Likely on TNE?

Having claimed that the naturalistic epistemologist cannot accept metaphysical naturalism without either inductive incoherence (if the probabilities of (1)–(5) above are high on neo-Darwinism) or skepticism (if the probabilities are unknown), Plantinga argues briefly that metaphysical supernaturalism—specifically, theism—escapes the difficulty. This is because, if we have been made by a theistic God, as creatures special to Him, then divine providence can be counted on to supply us with reliable faculties, cognitive and otherwise. So Plantinga:

[16] To this it should be added that one of the distinctive capacities of human beings is the ability to engage in linguistic communication. The adaptive advantages of this development can scarcely be overestimated. Yet linguistic communication, communication via a set of conventional symbols, is impossible unless the agents participating in the practice have beliefs which systematically correspond, on the whole, with what's true of the domain about which they are communicating, and only provided they generally say what they believe (as has been argued by Davidson and others).

> [The theist] has no corresponding reason for doubting that it is a purpose of our cognitive systems to produce true beliefs. . . . He may indeed endorse some form of evolution . . . guided and orchestrated by God. And *qua* traditional theist . . . he believes that God is the premier knower and has created us human beings in his image, an important part of which involves his endowing them with a reflection of his powers as a knower.[17]

Of course, having rejected Cartesian foundationalism, Plantinga thinks it would be circular to employ our reasoning faculties to argue *to* the existence of such a divine guarantor of our reasoning abilities. But he is prepared to claim that the belief that there is such a God and Maker of us, unlike metaphysical naturalism, supports a confidence in our cognitive powers. So the theist is not forced into incoherence or skepticism.

Unfortunately for TNE, Plantinga has no obvious right to this quick conclusion, unless he simply *builds it in* to his theistic hypothesis that God has created us as reliable knowers. The appeal to our having been created "in the image of God" will not secure that conclusion; for that could mean any number of things, and clearly we do not "reflect" God's nature as a knower very closely, or even remotely as closely as would be possible in creatures God could make.

We could, of course—and this is surely what Plantinga intends—understand our being created in God's image as implying that our cognitive faculties are reliable, at least in the domains in which they are properly employed. Would this help? Consider a parallel strategy on the part of the neo-Darwinian. Worried about Plantinga's argument, the neo-Darwinian might respond by tacking onto his theory some proposition such that, relative to the emended theory, the evolution of reliable belief-forming mechanisms becomes highly probable in organisms that make their living by reasoning. Whatever the details, there is every reason to suppose that the additional hypothesis, call it *H*, will not be inconsistent with standard neo-Darwinism (*D*).

Suppose that there is no significant independent evidence *for H* (nor against it), but that it is (roughly) probabilistically independent of *D*. (If it were probable relative to *D*, it would not be required as an independent hypothesis.) Now our evidence *R*—that human belief-forming mechanisms are quite reliable—is very probable relative to *D&H;* and so it appears that Plantinga's dilemma has been deflected. But in fact, the neo-Darwinian has conveniently ignored the fact that this maneuver comes at a price. In the absence of independent evidence for *H*, the prior probability of *D&H* is lower than that of *D* alone; and whether or not one gives this a Bayesian analysis, a clear constraint is that this prior must be *enough* lower that the probability of

[17] Plantinga, *Warrant and Proper Function*, 236.

R relative to naturalism cannot be raised by such a maneuver.[18] Failure to observe such a constraint would be to license cheating, by means of the most question-begging sort of ad-hoccery.

Let us turn now to theism. Traditionally, theism has been understood as the doctrine that there is just one God, creator of the universe, omnipotent, omniscient, perfectly good, and possessing essentially a variety of other perfections. But Plantinga does not understand what he calls theism in this way. Theism, as Plantinga takes it, includes all this but more as well. It includes, in particular, the view that God created humans in His own image. Perhaps it include a good deal else; perhaps Plantinga thinks you aren't a theist unless you hold that God's name is YHWH and that Moses was His prophet. However Plantinga constructs his version of theism, let's distinguish it from the traditional definition by calling the latter unvarnished theism (*UT*) and the former varnished theism (*VT*).

Now when it comes to showing that *R* is likely relative to theism, it is the doctrine that humans are created in the image of God (call this *I*) that does most of the work. So we can think of *VT* as just *UT*&*I*. As I shall show in a moment, *UT* cannot by itself be known to render *I* likely, so *I* is not a redundant part of *VT*. How much independent evidence is there, then, for *I*? So far as I can see, very little. The only generally recognized independent source for this doctrine is Gen. 1:26–27, which is echoed in various passages in the Old Testament and the New.[19] On the face of it, *I* makes a factual claim about an event (or series of events?) which occurred prehistorically.

[18] Plantinga, in a reply to a number of philosophers who have proposed this sort of defense of naturalism (he mentions Fred Dretske, Carl Ginet, Timothy O'Connor, Richard Otte, John Perry, Ernest Sosa, and Stephen Wykstra), makes just this point, though he poses the question in terms of the existence of defeaters for the warrant of a belief ("Naturalism Defeated").

[19] In the Old Testament, references to this idea outside of Genesis are very rare. The Hebrew term *tselem* is used mainly to refer to idols, that is, representations of foreign gods. Moreover, in Gen. 1:26–27 it is the "gods" (Elohim) who fashion man in *their* image—whatever that might mean. The New Testament passages further complicate matters. Rom. 8:29, II Cor. 3:12–4:4, and Col. 3:10 all imply that we have lost the divine image and need to recover it through Christ; I Cor. 15:49 goes so far as to suggest that we never had it and will not acquire it until we enter the Kingdom of Heaven. To make matters worse, I Cor. 11:7 appears to imply that men, but not women, were made in the divine image.

In any event, the Adamic fall creates a special difficulty for those theists, like Calvin, who hold that human beings have inherited from this event not merely moral or volitional depravity, but cognitive depravity as well. It is not clear how far this liability is supposed to go in depriving us of powers once possessed by Adam and Eve, but Calvin presumably did not believe—because it would have been foolhardy to do so—that grace or conversion to Christianity (or to Calvinism) somehow guarantees the restoration of the original human cognitive powers and faculties. Of course, it might seem entirely derelict on God's part for Him to have so arranged things that the consequences of a sin on the part of Adam and Eve could be transmitted to the rest of humanity. But then, what do we know?

Plantinga is more expansive in finding sources which warrant theism; he mentions Calvin's *sensus divinitatis*, Aquinas's natural but confused knowledge of God, the authority of the Church, and the internal testimony of the Holy Spirit in addition to Scripture as warrants for theism ("Naturalism Defeated"). Plantinga does not say which, if any of these, he takes to provide independent warrant for *I*. Needless to say, the theist will have to show what independent reasons

How seriously, then, should one take the testimony of Gen. 1:26–27? That topic truly requires more space than I can give it, so a few general observations will have to suffice. First, there is the familiar difficulty that the creation accounts in Gen. 1 and Gen. 2 are contradictory, and specifically, conflict with respect to the creation of human beings. Then, there is the generally mythical character of Genesis; many of the themes in the first eleven chapters are borrowed from, or influenced by, the myths of other ancient Near Eastern cultures. Further, there is the archaeological evidence that the story of the Egyptian captivity and Exodus are mythical, not historical accounts.[20] All these considerations undermine the reliability of Genesis as a source of historical information. Finally, there are manifold problems of interpretation: just what did the author of Gen. 1 mean in saying that we were created in God's image? Did that author actually have our cognitive faculties in mind?

One might also construct an argument of sorts for the desired conclusion from natural theology. Such an argument would move from personal attributes of God, including, for example, His free will, to the conclusion that we, who are also persons, are created in His Image. (In that case, *I* would *not* be independent of *UT*—at least not if it could be shown that God's goodness makes it probable that He would create personal beings.) But the trouble with this argument—as with reliance on Genesis—is that it says nothing about how *good* an image we are.[21] Indeed, in view of considerations I shall now adduce, the argument *cannot* hope to do this.

I said above that *I* does most of the work in securing the likelihood of *R* on *VT*; that is, just as God's cognitive faculties are perfect, so we, as imperfect images of God, can count on our cognitive faculties being good ones in their limited way. But unless *R* is actually built into *I*, we have no right to conclude even this.

The difficulty is this. Even if God has the traditional theistic attributes and made us in His image, we have no right to draw conclusions about what God would do or would be likely to do, in making us, unless we had a much firmer grasp on God's purposes and judgments concerning what is good and what is

---

there are for the reliability—or even the existence—of these alleged sources. Of course, the theist who believes that God has *told* her that her cognitive faculties are in good shape can't be faulted on the score of having (in the way at issue) an incoherent or self-undermining set of beliefs; but neither can the schizophrenic who believes that God has told him that he is the *parousia*. I discuss the epistemic status of revelations in greater detail in Fales, "Mystical Experience as Evidence," *International Journal for Philosophy of Religion* 40 (1996): 19–46; "Scientific Explanations of Mystical Experience, Part I: The Case of St. Teresa," *Religious Studies* 32 (1996): 143–163; and "Scientific Explanations of Mystical Experience, Part II: The Challenge to Theism," *Religious Studies* 32 (1996): 297–313. Plantinga himself addresses these issues in *Warranted Christian Belief* (Oxford: Oxford University Press, 2000).

[20] For details see Hershel Shanks et al., *The Rise of Ancient Israel* (Washington, D.C.: Biblical Archeology Society, 1992).

[21] A second difficulty, if the argument trades simply on the fact that we, like God, are persons, is that the New Testament passages cited in note 19 can hardly be made sense of on the view that possessing the Image is merely a matter of being a person.

not, than we evidently do. Even if theists have no independent reason to suspect their cognitive faculties of serious fallibility in more ordinary contexts, *this* context gives them eminently good reasons for doubt.

The problem of evil is itself sufficient to raise this doubt. Theists can argue that the logical incompatibility of observed evil with theism hasn't been demonstrated. But unless they can show that observed evil is *likely* given theism, their only defense will be to claim that we are radically ignorant of God's purposes—too ignorant even to conceive what sort of justifying reasons God has for permitting some horrific evils. But then, it appears that we are not really very good knowers, when it comes to grasping the divine scheme of things.[22] Indeed, *one* of the apparent evils which we cannot explain (as Descartes saw) is why God made us such poor knowers, so fallible and so unable to understand Him and His moral order better. Perhaps—for some reason we doubtless can't fathom—God has even made us very poor at natural theology and at receiving and interpreting His revelations.[23] That would be unsurprising, on the supposition that God saw fit to make us defective in these ways.

Given all this, the theist has no right to conclude that, relative to *VT*, her cognitive faculties are as good as she would like to think they are. If God can see fit to allow small children to die of terrible diseases for some greater good we cannot imagine, might He not have given us radically defective cognitive systems, and allowed us to be lulled into thinking them largely reliable, also for some unimaginable reason? Plantinga is quite correct in observing that *VT* provides no reason to think our cognitive reliability is low. But it also gives no reason for thinking it is high—and that is the claim Plantinga needs. It is the theist, not the metaphysical naturalist, who is in trouble here.

But suppose now that theists do interpret *I* in such a way that *R* is rendered likely by *VT*. Would that help? Clearly, the theist now has an escape from the loop of skeptical doubts—if she has good reason to think that theism (in this sense) is true. Recall, however, the naturalist. He, too, can escape the skeptical loop by adding *H* to *D*—*if* he has good independent grounds for thinking that *D*&*H* is true. But his grounds for thinking that were (we supposed) no better than his grounds for *D* alone. So if *D* alone gives no good reason for thinking that *R* is true, *D*&*H* can hardly do so—*H* will be as unlikely on the available evidence as *R* itself is. But the same applies to the theist's *I*, as now understood. If *R* has a low or unknown likelihood on *UT*, then, because *I* (as now understood) makes *R* likely, and because the theist has precious little in-

[22] Plantinga admits as much in "Self-Profile," in *Alvin Plantinga*, ed. James E. Tomberlin and Peter Van Inwagen (Boston: D. Reidel, 1985); and "Epistemic Probability and Evil," in *The Evidential Argument from Evil*, ed. Daniel Howard-Snyder (Bloomington: Indiana University Press, 1996).

[23] We have abundant evidence of that (if evidence counts for anything under these circumstances) in the very small number of revelations, out of all the purported ones, that could possibly be true, given the conflicts between them, and in the significant number of falsified Biblical prophecies.

dependent evidence for *I*, it will follow that *I*—and hence *UT*&*I*—has a low or unknown likelihood, relative to *UT* plus all the evidence. But then, even if the theist has good reasons for believing *UT*, she has no good reason for believing *VT* under the present interpretation of *I*. And that means that she, too, has no way of escaping the skeptical loop.

IV. Human Fallibility

I haven't yet explained how the findings in cognitive psychology alluded to above are to be explained. But it should be clear, first, that they are equally an embarrassment for the theist. Descartes was concerned to show that our cognitive faculties were so designed by a beneficent God that, properly employed, they would not deceive. But the fallacy-generating mechanisms uncovered by psychologists suggest a different story. For most of us, in the relevant circumstances, the correct inferences are cognitively unavailable, even upon reflection. Where they are available, the impulse to reason incorrectly still appears to be strong and innate. Why?

Neo-Darwinians have several natural ways to answer that question. In part, they can use the very strategies Plantinga has charged against them. They can point out (1) that evolution is opportunistic: it can only work with the materials at hand, with lucky mutations, and with selective pressures that are common and persistent; and (2) that evolution can permit some degree of maladaptiveness, especially when associated benefits outweigh costs. To use Herbert Simon's term, evolution is a satisficer; and so are we. How might Darwinians marshal these two points?

When it comes to details, it is evident that no single explanatory strategy can apply to every cognitive limitation; Darwinians will have to proceed on a case-by-case basis. But some illustrative remarks will serve to indicate the procedure. I shall discuss briefly two types of cognitive failure: visual illusions and inductive inferences to beliefs about cause-effect relationships.

Visual illusions typically achieve their effects by taking advantage of such visual signal-processing mechanisms as edge detection, etc., which under normal circumstances are essential to our being able to make visual sense of our surroundings. Illusions trick these mechanisms by co-opting the very information-processing algorithms whose automatic operation serves us faithfully in the environment for which they are adapted. Take size constancy. When two one-inch human figures are drawn alongside railroad tracks receding into the distance, the "nearer" figure looks larger; the perspectivally drawn tracks trigger the size-constancy mechanism. But that's just as it should be in real life, where it's more important to gauge the size and distance of an approaching saber-toothed cat than to interpret a psychologist's drawing.

Causal judgments provide an example from the sphere of inferential processes. As Hume supposed, we commonly make such judgments by notic-

ing correlations between instances of one type of event, *C*, and instances of another, *E*. But humans often use the wrong procedure in estimating the statistical relevance of *C* to *E* by tending to ignore the frequency with which *E* occurs when *C* has not occurred. And this can lead to mistaken causal judgments.[24]

Why are we so addicted to an incorrect procedure for estimating causal relevance, since incorrect causal judgments can sometimes be costly? A Darwinian answer, I suggest, will rely upon two points. First, on an evolutionary time scale, the invention of arithmetic is extremely recent; computation of the relevancy of *C* to *E* would, for most of our biological existence, have been carried out by less explicit processes. And that may not be such an easy task. Computing in one's head such a simple function as the difference between two fractions is hard even *with* arithmetic. Second, it may ordinarily not have mattered much: for most purposes, a rough estimate of $P(E/C)$ is good enough. More than that: estimating $P(E/C) - P(E/\sim C)$ is far more time-consuming, and in many cases, survival may have depended upon the *rapid* making of causal judgments. Thus, and in general, the costs—temporal and otherwise—of accurate, loophole-free reasoning must be weighed against the benefits of improved reliability.[25] Evolution tends to home in on organisms that can get things right and get things fast, sacrificing only as much of each for the other as necessary for effective action. For creatures such as humans who have found their ecological niche by specializing heavily in reasoning and the flexibility that a powerful capacity to learn affords, this conclusion applies in spades—certainly for the handling of more mundane beliefs.[26]

Finally, how should we estimate the reliability of *theoretical* inferences, especially those in science, given neo-Darwinism? That is too large a subject to address here, but in a nutshell, the answer is surely, first, that such inferences are indeed somewhat more risky; and second, that they rely very largely on reflective refinements of the same computational strategies that have been tested and improved by eons of evolutionary development for the ordinary purposes of survival and procreation. Many of *those* inferences are also quite theoretical: witness the young swain's earnest attempts to guess what his intended is thinking.

[24] See Kahneman, Slovic, and Tversky, *Judgment under Uncertainty*, esp. chaps. 8, 10. For a further discussion of these findings and a survey of the literature, see Fales and Edward A. Wasserman, "Causal Knowledge: What Can Psychology Teach Philosophers?" *Journal of Mind and Behavior* 13 (1992): 1–27.

[25] A discussion of these trade-offs is provided by Goldman, *Epistemology and Cognition*, especially in chapters 6 and 14.

[26] The evidence suggests that humans have evolved two sorts of reasoning capacities: slow, flexible ones that respect correct and universal rules of inference, and a set of quick-and-dirty algorithms of narrower scope that yield approximate truth most (or enough) of the time.

PART II

SKEPTICISM

Commonsense Naturalism

Michael Bergmann

I. Introduction

Metaphysical naturalism is, roughly speaking, the view that there are no supernatural beings—no such beings as, for example, God or angels or ghosts.[1] Thomas Reid was a theist and, therefore, not a naturalist. Consequently, one wouldn't expect to find in Reid's writings an argument in support

Thanks to Jan Cover, Keith Lehrer, Trenton Merricks, Michael Rea, William Rowe, Dale Tuggy, and especially Thomas Crisp and Alvin Plantinga for comments on earlier drafts. An earlier version of this chapter was read at the Twentieth World Congress in Philosophy in Boston. My thanks to the audience members as well as my fellow presenters, Richard Otte and William Ramsey, for helpful advice. Finally, I would like to acknowledge the support of the Purdue Research Foundation for a Summer Faculty Grant that enabled me to work on this project.

[1] This is how Alvin Plantinga characterizes the view he attacks in his evolutionary argument against naturalism ("Naturalism Defeated," unpublished). See also Plantinga, "Respondeo," in *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, ed. Jonathan Kvanvig (Lanham, Md.: Rowman and Littlefield, 1996), 350–352. Since the naturalism under discussion in this paper is the sort—whatever it is—that Plantinga is attacking, it would be best to begin with his understanding of it. Another similar characterization of metaphysical naturalism is Michael Devitt's. He says that it amounts to physicalism—the view that all entities are physical entities ("Naturalism and the A Priori," *Philosophical Studies* 92 [1998]: 46). If I were trying to give a precise account of metaphysical naturalism, much more would have to be said. But this will do for our purposes. See the essays in *Midwest Studies in Philosophy*, vol. 19, *Philosophical Naturalism*, ed. Peter French, Theodore Uehling, and Howard Wettstein (Notre Dame: University of Notre Dame Press, 1994), and in *Naturalism: A Critical Appraisal*, ed. Steven Wagner and Richard Warner (Notre Dame: University of Notre Dame Press, 1993), for various attempts to clarify what naturalism is, as well as the essays in *Objections to Physicalism*, ed. Howard Robinson (New York: Clarendon Press, 1993), for discussions of what physicalism is. See also section 7 of this chapter, where I discuss varieties of naturalism other than the sort roughly defined here. For a defense of the view that naturalism isn't a thesis at all but rather a research program (more specifically, a plan to conduct inquiry using only the methods of the natural sciences), see

of naturalism. But one *can* find in Reid the resources for a defense of naturalism against a certain sort of objection to it. In this chapter I will propose a Reid-inspired commonsense response to Alvin Plantinga's evolutionary argument against naturalism. It is a response whose relevance extends far beyond Plantinga's argument. For it also serves as a preliminary defense and illustration of some of the main elements in a commonsense response to skepticism.

Plantinga has recently argued (in *Warrant and Proper Function*, chapter 12, and in "Naturalism Defeated")[2] that naturalism is self-defeating. He asks us to imagine a race of creatures about whom we know nothing except that they form and change beliefs and that they came into existence via the mechanisms of evolution. Then he asks us to consider the probability that the cognitive faculties of these creatures are reliable—more specifically, he asks us to consider the probability that their cognitive faculties are reliable given naturalism and evolution. We can express this probability as P(R/N&E) where 'R' is the claim that the cognitive faculties in question are reliable, 'N' is the claim that naturalism is true, and 'E' is the claim that these faculties came into existence by way of the mechanisms of evolution. Plantinga thinks P(R/N&E) is low or inscrutable because evolutionary processes aim at adaptive behavior and having reliable faculties doesn't seem particularly probable with respect to adaptive behavior. This is so, he thinks, when P(R/N&E) is specified to the hypothetical creatures mentioned. But he also thinks P(R/N&E) is low or inscrutable when we specify it to ourselves—there being no relevant difference between ourselves and the creatures in his example.

That's the first stage of Plantinga's argument. In the second stage he points out that the fact that P(R/N&E) is low or inscrutable constitutes a *defeater* for R for anyone who endorses N&E. Then he says that if you're a naturalist, the sensible thing for you to believe is that evolution is true (you have no recourse to divine creation). So the naturalist should believe N&E. But then, once apprised of Plantinga's argument, the naturalist will have a defeater for R. And a defeater for R is a defeater for every one of a person's beliefs—including belief in N. This, says Plantinga, makes naturalism self-defeating. (Notice that Plantinga's argument can be construed as an argument—starting from naturalistic premises—for global skepticism. This is why my Reidian response to it can be used as an example of how to respond to more typical skeptical challenges.)

For the purposes of this chapter, I will grant to Plantinga the conclusion of the first stage of his argument—that P(R/N&E) is low or inscrutable when specified to us. My contention is simply that this does not necessarily constitute a defeater for R (for the supporter of N&E). In order to defend this view

---

Michael Rea, *World without Design: The Ontological Consequences of Naturalism* (Oxford: Clarendon Press, 2001).

[2] *Warrant and Proper Function* (New York: Oxford University Press, 1993), hereafter *WPF*; and "Naturalism Defeated."

I will first explain, in the next section, a response Plantinga gives to the probabilistic argument from evil. Then, in sections 3 and 4, I will present a view of Reid's that makes possible a *Reidian* response to Plantinga's evolutionary argument against naturalism that parallels *Plantinga's* response to the probabilistic argument from evil.[3] In sections 5–8 I will develop Reid's commonsense response in the context of considering a variety of objections to it. I will conclude in section 9 by connecting that response as well as my defense and development of it with the more general issue of skepticism.

I should note at the outset that my Reidian response isn't merely an ad hominem attack on Plantinga. True, Plantinga endorses a Reidian epistemology, so a Reidian response patterned after a response Plantinga himself gives in another setting will, if successful, create a special problem for him. But the Reidian response I offer relies on elements of Reid's epistemology that have a much wider appeal than does Plantinga's own epistemology. It depends on the Reidian views that (i) a belief can be *noninferentially* justified or warranted—that is, justified or warranted even if formed on the basis of an experience rather than on the basis of another belief[4]—and that (ii) among our noninferentially justified beliefs are a good number of our commonsense beliefs. The sort of foundationalism inherent in (i) is not the least bit unusual among contemporary epistemologists. And the commonsensism endorsed in (ii) is thoroughly intertwined with the particularist approach to philosophical analysis that is commonly employed in contemporary metaphysics, ethics, and epistemology.[5] So although the response I propose will be of no use to those who reject (i) and (ii), its benefits are by no means limited to those who accept Plantinga's epistemology.

One more preliminary remark. Some will wonder if my use of Reid (a the-

[3] Keith Lehrer's response to Plantinga's evolutionary argument against naturalism is superficially similar to my own. In giving his response to that argument he too draws upon Plantinga's response to the problem of evil (see Lehrer, "Proper Function versus Systematic Coherence," *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, ed. Jonathan Kvanvig [Lanham, Md.: Rowman and Littlefield, 1996]) and upon Reid (see Keith Lehrer and Bradley Warner, "Reid, God and Epistemology," *American Catholic Philosophical Quarterly* 74 [2000]: 357–372). But the use Lehrer makes of Reid and of Plantinga's response to the problem of evil is quite unlike the use I make of them. So our responses to Plantinga's evolutionary argument against naturalism are, in the end, very different (which is what one might expect given that I'm an externalist foundationalist and that Lehrer is at least very sympathetic to an internalist sort of coherentism).

[4] In saying that (i) is a Reidian view, I'm assuming that he is a foundationalist. Lehrer challenges that assumption. In support of a coherentist reading, Lehrer points out passages in Reid which could be taken as saying that the justification of each of our beliefs depends on a further belief about the trustworthiness of that original belief's source. See Lehrer, "Chisholm, Reid and the Problem of the Epistemic Surd," *Philosophical Studies* 60 (1990): 42–43. This isn't the place to defend the view that Reid is a foundationalist. So I'll just say that I read those passages in Reid as saying that we *can* be justified in the further belief that our belief sources are trustworthy, not that we *must* be in order for our beliefs produced by those sources to be justified.

[5] See Roderick Chisholm, "The Problem of the Criterion," in *The Foundations of Knowing* (Minneapolis: University of Minnesota Press, 1982), 61–75, for an account of this particularist approach.

ist) in defense of naturalism is something of which Reid himself would approve. To soften up such readers, I will include a quotation from Reid in which he suggests that one needn't be a theist to believe with justification in the reliability of one's senses:

> Shall we say, then, that this belief [in the reliability of our senses] is the inspiration of the Almighty? I think this may be said in a good sense; for I take it to be the immediate effect of our constitution, which is the work of the Almighty. But, if inspiration be understood to imply a persuasion of its coming from God, our belief of the objects of sense is not inspiration; for a man would believe his senses though he had no notion of a Deity. He who is persuaded that he is the workmanship of God, and that it is a part of his constitution to believe his senses, may think that a good reason to confirm his belief. But he had the belief before he could give this or any other reason for it.[6]

I'm sure that Reid would say that similar remarks apply to the naturalist's belief in R.

II. Plantinga on the Argument from Evil

Plantinga has a lot to say about the probabilistic argument from evil.[7] I don't propose to discuss all of it here. But one thing he says is of particular interest for our purposes. Suppose that P(G/HE) is low (where G is the claim that God exists and HE is the claim that there are horrendous evils). What follows concerning the rationality or reasonableness or warrant for the belief that G? Not much, says Plantinga. For someone who believes that P(G/HE) is low might also believe some other proposition Q and recognize that P(G/HE&Q) is high. If so, the fact that she also believes that P(G/HE) is low won't make it unreasonable for her to believe G. But, says Plantinga, suppose we grant to the atheist objector that P(G/k) is low (where k is the *total* relevant propositional evidence at one's disposal). What follows then concerning the rationality of holding G? Again not much, says Plantinga.

Here's an example he uses to explain why.[8] Suppose that a letter has gone

[6] Thomas Reid, *Essays on the Intellectual Powers* (Cambridge: MIT Press, 1969), 294–295; hereafter Reid, *Essays*. See also Reid, *Inquiries and Essays*, ed. Ronald Beanblossom and Keith Lehrer (Indianapolis: Hackett Publishing, 1983), 203; hereafter Reid, *Inquiries*. Reid's *Essays* is complete but his *Inquiries* (which isn't complete) is more readily available. I will give references to both where possible.

[7] See Plantinga, "The Probabilistic Argument from Evil," *Philosophical Studies* 35 (1979): 1–53; "Epistemic Probability and Evil," *Archivo di filosofia* 56 (1988); "On Being Evidentially Challenged," in *The Evidential Argument from Evil*, ed. Daniel Howard-Snyder (Bloomington: Indiana University Press, 1996), 244–261; and "Degenerate Evidence and Rowe's New Evidential Argument from Evil," *Noûs* 32 (1998): 531–544.

[8] Plantinga, "Epistemic Probability and Evil," in *The Evidential Argument from Evil*, ed. Daniel Howard-Snyder (Bloomington: Indiana University Press, 1996), 88–89.

missing, that you have an obvious motive for stealing it, and that both circumstantial evidence and eyewitnesses place you at the scene of the crime with ample opportunity to steal the letter. You claim to have been out alone for a walk in the woods at the time the letter was stolen (call this claim 'W'). But because of the strength of the case against you (and the fact that you have done things of this sort in the past), others are extremely doubtful of W. They sensibly conclude that P(W/k)—where k is their total propositional evidence—is quite low. However, you clearly remember being out in the woods for a walk earlier in the day at the time the letter was stolen. This memory involves a belief ground that is *nonpropositional;* it involves a *seeming* of some sort that results in your taking a particular memory belief to be obviously correct under the circumstances (the phenomenology of these belief grounds is familiar enough but it is very difficult to describe). You, unlike those who think you are guilty, have the experiential evidence of *its seeming to you like you were out for a walk earlier in the day* and that very evidence grounds the belief that you did not steal the letter.[9]

So you know you didn't steal the letter and you know this on the basis of *non*propositional evidence. Nevertheless, your total relevant *propositional* evidence is more or less the same as that of those who think you are guilty.[10] You too agree that P(W/k) is low. Yet this doesn't in the least suggest that you are irrational to believe W; for you clearly remember being out in the woods at the time in question. The point is that a proposition's being improbable on everything else you know or believe doesn't make belief in it irrational. And this is so even if these other things you believe are clearly *relevant* bits of evidence. For you may have in addition to all the *propositional* evidence at your disposal certain *non*propositional evidence. And this nonpropositional evidence may be strong enough to make it completely reasonable for you to hold the belief in question even while recognizing that the belief is improbable on your total relevant *propositional* evidence. Furthermore, given that this total relevant propositional evidence is all your accusers have to go on, you can also concede that your accusers are completely reasonable in thinking you are guilty. They are in this unfortunate situation because they lack an important bit of (nonpropositional) evidence that only you have.

Plantinga applies these considerations to the theist confronted with the

[9] When I describe the experience as one of its *seeming to you that you were out for a walk in the woods earlier in the day*, I don't mean to suggest that the experience has a propositional content. I'm just saying that the experience in question is a seeming that inclines you to believe that you were out for a walk in the woods earlier in the day.

[10] We have to imagine the case so that you have all the propositional evidence your accusers have and that additionally all you have is the memory experience and the belief that W. In particular, you don't have any beliefs about your memory experience or about how trustworthy it is, etc. Your belief that W is based solely on the memory experience in question and nothing other than the belief that W is based on that memory experience. One might think that, upon being accused and thinking carefully about your memory experience, you will form additional beliefs on the basis of it. But we can stipulate that we are focusing on the time before you are accused—the point at which you first learn of all the evidence that exists against you.

probabilistic argument from evil in the following way. A person might have sufficiently strong nonpropositional evidence for G by way of what John Calvin calls 'the *sensus divinitatis*'. This faculty triggers belief in God (or beliefs about God) in response to certain experiences and circumstances. We observe the beauty and majesty of a starry night, are overwhelmed with a sense of awe, and find ourselves thinking *God has created this universe;* we recognize that we have done something that is wrong, feel guilty before God, and find ourselves thinking *God disapproves of this;* when life is sweet and satisfying we are overcome with a sense of gratitude and believe *God is to be thanked and praised*. In each case, we have a belief about God formed not on the basis of other beliefs but on the basis of experiences.[11] In cases where this evidence is sufficiently strong, one can come to rationally believe in God's existence despite recognizing that P(G/k) is low. The nonpropositional evidence makes rational a belief that is improbable with respect to one's total relevant propositional evidence.

In evaluating Plantinga's response to the argument from evil, one might wonder whether there *is* any such nonpropositional evidence for theism and, if so, how strong it is. But the main point I want to draw attention to is that the belief that P(G/k) is low does not in itself constitute a defeater for G (for the person whose total relevant propositional evidence is k). In addition, it must be the case that the person in question has no sufficiently strong *non*-propositional evidence for G.

III. Reid on Knowing R

Reid (or Reid as I understand him) says that we know R not by basing that belief on other beliefs but instead in the basic way. According to Reid, R is a first principle: "Another first principle is, that the natural faculties, by which we distinguish truth from error, are not fallacious."[12] And first principles, says Reid, are properly believed *noninferentially*.[13] We obtain this noninferential knowledge of first principles—which he also calls 'principles of common sense', 'self-evident truths', and 'intuitive judgments'—by employing that branch of our faculty of reason he calls 'common sense'.[14] The idea isn't that we have a faculty for knowing in the basic way things like "You shouldn't try to drive downtown in a hurry during rush hour." Common wisdom of this latter sort is cultural and learned and will vary across times and places. Reid is

[11] Plantinga, "Reason and Belief in God," in *Faith and Rationality*, ed. Alvin Plantinga and Nicholas Wolterstorff (Notre Dame: University of Notre Dame Press, 1983), 80–81.

[12] Reid, *Essays*, 630; Reid, *Inquiries*, 275.

[13] Reid, *Essays*, 593.

[14] Ibid., 567. The other branch of reason enables us "to draw conclusions that are not self-evident from those that are."

talking about something else. He's speaking of a faculty whereby we form beliefs naturally held by sane humans in normal circumstances—noninferential beliefs that are not the result of education but of our constitution (though they are certainly acquired sometime after birth).[15] Reid thinks that by means of this faculty we know both contingent and necessary truths.[16] What he thinks of as knowledge via common sense of necessary truths is what we would call 'a priori knowledge'. Examples he gives of necessary truths known via common sense are the axioms of logic and mathematics.[17] Examples he gives (in addition to R) of contingent truths known noninferentially via common sense are beliefs such as "The thoughts of which I am conscious are *my* thoughts," "Other humans have minds," and "I have some degree of control over my actions."[18]

So Reid thinks we know R and other first principles (both contingent and necessary) in the basic way by means of common sense. Now, just as there is a mechanism by which we form sense perceptual beliefs in the basic way, so also there is a mechanism of sorts for forming our commonsense beliefs. Sense perception seems to work as follows: we experience sensations (visual, tactile, etc.) and on the basis of them form beliefs in the existence of external objects having certain qualities. The ground of our sense perceptual beliefs is our sense experience, not other beliefs. It is because they aren't based on other beliefs that they are called basic or noninferential. Now consider what Reid says about how commonsense beliefs in first principles are formed:

> We may observe, that opinions which contradict first principles are distinguished from other errors by this; that they are not only false, but absurd: and, to discountenance absurdity, nature has given us a particular emotion, to wit, that of ridicule, which seems intended for this very purpose of putting out of countenance what is absurd, either in opinion or practice.[19]

The idea is that when we entertain the contrary of a first principle, we experience the emotion of ridicule. On the basis of this experience, we dismiss as absurd the contrary of the first principle and believe the first principle. In other words, we consider the contrary of a first principle and have an experience that prompts this sort of belief: "That's absolutely nuts! It's ridiculous!" It thereby also prompts belief in the first principle itself, though, as Reid notes, we rarely attend to beliefs in first principles.[20] Just as in the case of

[15] See ibid., essay 6, chap. 4, "Of First Principles in General," parts of which are included in Reid, *Inquiries*.
[16] Reid, *Essays*, 614–615.
[17] Ibid., 644; Reid, *Inquiries*, 284–285.
[18] Reid, *Essays*, 611–643.
[19] Ibid., 606; Reid, *Inquiries*, 259.
[20] Reid, *Essays*, 632–633; Reid, *Inquiries*, 277.

sense perception, the ground of the first principle belief is an experience not a belief.

IV. A Reidian Response to Plantinga

It should now be pretty obvious how a Reidian could respond to Plantinga's evolutionary argument against naturalism. She could combine *Plantinga's* method of responding to the probabilistic argument from evil with *Reid's* account of how we can know R in the basic way. For the sake of argument, we've conceded to Plantinga that P(R/N&E) is low or inscrutable. He says this is a defeater for R. But the commonsense naturalist can respond as follows: "Even if a naturalist believed that P(R/N&E) is low or inscrutable, this needn't give her a defeater for R. For she could have *non*propositional evidence for R that is sufficiently strong to make belief in R rational, reasonable, and warranted—even for someone whose total relevant *propositional* evidence, k, was such that P(R/k) is low or inscrutable. The nonpropositional evidence she has could be of the sort Reid describes."

To clarify this Reidian response, let me briefly consider two objections to it that are based on misunderstanding. The first has to do with the parallel between ourselves and the hypothetical creatures mentioned in stage one of Plantinga's evolutionary argument. It seems that the conjunction of the belief that N&E and the belief that P(R/N&E) is low or inscrutable *does*—when specified to these hypothetical creatures—constitute a defeater for our belief that *their* cognitive faculties are reliable. But then why should the same beliefs specified to ourselves not constitute a defeater for R specified to us? There is no relevant difference between the two cases since the facts concerning our origins are the same.

But there is a relevant difference. It may be true that if P(R/N&E) is assigned a low value when specified to the hypothetical creatures then it should also be assigned a low value when specified to us. But it's *not* true that if belief in that low probability claim results in a defeater in the case of the hypothetical creatures it also results in a defeater in our own case. First let's be clear about what exactly gets defeated in the case of the hypothetical creatures. It is *our* belief that *their* faculties are reliable. But notice that, in thinking about these hypothetical creatures, all we have to go on is *propositional* evidence; we have no *non*propositional evidence for R specified to them (recall how little we know about them). That's why it is plausible to think that the belief that N&E along with the belief that P(R/N&E) is low or inscrutable (where both beliefs are specified to these hypothetical creatures) constitutes a defeater for our belief that *their* faculties our reliable. But of course things are different with our belief in the reliability of *our own* cognitive faculties. In our own case, we have *non*propositional evidence *in addition to* the sort of propositional evidence we have in the case of the hypothetical creatures. That's why the belief that P(R/N&E) is low or inscrutable along with the belief that N&E

(where both beliefs are specified to us) does not constitute a defeater for our belief that our own faculties are reliable.[21]

A second objection is that the Reidian response implies that R is beyond defeat. But R *could* be defeated. Suppose someone became convinced that she was the victim of a Cartesian demon. This would give her a defeater for R.

That seems right. But nothing I've said conflicts with it. Consider again the example of your being falsely accused of stealing a letter when you clearly remember your innocence. The circumstantial (propositional) evidence fails to defeat your memory belief. But that doesn't mean that your memory belief is beyond defeat. You could become convinced that the memory in question was planted in you artificially by someone intending to deceive you. This would create a defeater for it. Or consider theism and horrendous evil. You might believe G in the basic way and thereby have a lot of warrant for it. If so, then the fact that you also think P(G/HE) is low does not defeat your belief that G. But you could become convinced that your belief in G is the product of a Freudian sort of wish fulfillment—a way of forming beliefs that you take to be unreliable. Then you would have a defeater for your theistic belief.[22] In the same way, belief in the Cartesian demon might be a defeater for R even though belief in N&E together with belief in the low probability of R on N&E is not. Notice that the presence of these defeaters (the Freudian defeater for G or the Cartesian defeater for R) is compatible with the existence of nonpropositional evidence for G or for R. It's not that this nonpropositional evidence has no effect; rather, it's that its effect has been defeated by the stronger contrary effect of the defeater in question.[23]

V. A Disanalogy?

I've claimed that my Reidian response to Plantinga's evolutionary argument against naturalism is analogous to Plantinga's response to the probabilistic argument from evil. But is it? Here's a reason to think not. The ex-

[21] Thus Plantinga is right when he says in *WPF*, 229, that the person considering R specified to the hypothetical creatures has no source of information about R other than the propositional evidence mentioned. But when he considers (233–234) what other sources of information we might have for R specified to *us*, he considers only other *propositional* evidence for R. And he considers it only as a candidate for being a defeater of a defeater for R instead of thinking of the other source of information about R as something that prevents us from having a defeater for R in the first place. He fails to acknowledge that we have nonpropositional evidence for R specified to us and that this is a relevant difference between the two cases; it's a difference that results in our having a defeater in the one case and not in the other.

[22] Notice that the mere *existence* of the Freudian explanation is not in itself a defeater for G, just as the mere existence of the Cartesian demon hypothesis is not in itself a defeater for R. It must also be the case that the alternative hypothesis in question is reasonable and/or believed.

[23] Those sane humans for whom R is defeated (assuming there are such) are not counterexamples to my earlier suggestion that the outputs of R are beliefs naturally held by *all* sane humans in normal circumstances. For in order to be defeated, these beliefs had to first be held. And, in fact, they *are* held in normal circumstances; it is only in *abnormal* circumstances that someone comes to later believe she is the victim of a Cartesian demon.

ample of the Freudian defeater for theism mentioned in the previous section is instructive. For it is successful in the face of nonpropositional evidence for theism even though the belief that P(G/HE) is low, used in the argument from evil, is not. The reason for this, one might think, is that the Freudian defeater is an *undercutting* defeater. It works against theistic belief by casting doubt on the trustworthiness of its *source*. In this way it is unlike the argument from evil, which is a *rebutting* defeater—a defeater that provides evidence for the *falsity* of theistic belief. If a person has sufficiently strong nonpropositional evidence for her theistic belief, it will not be defeated by her belief in the potential rebutting defeater that the probability of theism on her propositional evidence is low. Instead, that person will learn merely that one of her beliefs—her theistic belief, which she still reasonably holds—happens to be unlikely on her propositional evidence. But when she comes to have doubts about the trustworthiness of the source of her theistic belief, then she does have a defeater—an undercutting one. The nonpropositional evidence for theism doesn't help because its value as evidence is now in doubt.

This helps us to see how the Reidian response to Plantinga's proposed defeater for naturalism is supposed to be disanalogous to Plantinga's response to the defeater for theism from horrendous evil. Plantinga's response to the problem of evil worked (we'll waive concerns about whether theistic belief is properly basic) because the defeater from horrendous evil didn't cast doubt on the *source* of theistic belief. But, one might argue, the Reidian response won't work because Plantinga's proposed defeater for naturalism *does* cast doubt on the source of our beliefs—including the belief in naturalism. This is an important point. For in the previous section I conceded that the Freudian defeater for theistic belief and the artificial-memory-implant defeater for one's memory beliefs can be successful. What these defeaters have in common is that they are undercutting defeaters that focus on the source of the belief to be defeated with the aim of casting doubt on that source's trustworthiness. If I grant that such defeaters can work, how can I resist Plantinga's proposed defeater for naturalism, which works in the same way?

In order to answer to this question, it will be helpful to be more precise about what the Freudian defeater, the artificial-memory-implant defeater, and the Plantingian defeater for naturalism have in common. Here's the structure they share.[24] We'll call the belief threatened with defeat 'B'. The proposed defeater for B consists of two further beliefs:

X: The source of B is of kind K.
Y: P(a belief source S is reliable/S is of kind K) is low or inscrutable.

[24] *If* they share a structure at all. See below where I raise doubts about whether the Freudian and memory implant defeaters really involve probabilistic inference in the way that Plantinga's proposed defeater does.

This pair of beliefs is supposed to result in a defeater for B because a person who believes both X and Y will come to have grave doubts about the truth of

Z: The source of B is reliable.

And doubts about Z are the very sort of thing that result in an undercutting defeater for B. (Notice, by the way, that one doesn't need to actually believe that Z is false in order for B to be thus defeated.)[25] We can see as follows that each of the three defeaters under consideration is of the above form.

Freudian Defeater for Theism
Because of the following beliefs:
X: The source of my belief that G (i.e., that God exists) is wish fulfillment
Y: P(a belief source is reliable/it is a sort of wish fulfillment) is low or inscrutable
it is reasonable to doubt:
Z: The source of my belief that G is reliable.

Artificial-Memory-Implant Defeater for W
Because of the following beliefs:
X: The source of my belief that W (i.e., that I was out for a walk in the woods at the time the letter was stolen) is artificial input by scientists who, as far as I know, have no particular interest in the beliefs so produced being true
Y: P(a belief source is reliable/it consists of input by scientists who, as far as I know, have no particular interest in the beliefs so produced being true) is low or inscrutable
it is reasonable to doubt:
Z: The source of my belief that W is reliable.

Plantinga's N&E Defeater for Naturalism
Because of the following beliefs:
X: The sources of our beliefs (including my belief in naturalism) came about by way of N&E
Y: P(a belief source is reliable/it came about by way of N&E) is low or inscrutable
it is reasonable to doubt:
Z: Our belief sources are reliable (i.e., R).

By way of contrast, consider the form of the defeater for theism from evil:

Defeater for Theism from Evil
Because of the following beliefs:
X: HE (i.e., there exists horrendous evil)
Y: P(G/HE) is low or inscrutable

[25] It's also worth noting that to agree that doubts about Z are sufficient to defeat B is not to agree that belief in Z is necessary for B's justification or warrant.

it is reasonable to doubt:
B: G.

In each of these four cases, the person holding the threatened belief B (whether B is theism or W or naturalism) is supposed to have a defeater for it upon learning that the relevant X and Y are true. But in the first three cases, the defeater is an undercutting one that works by causing doubts about Z—the belief about the *source* of B—whereas in the last case, the defeater is a rebutting one that works by casting doubt on the truth of B itself.[26] What is supposed to cause problems for my Reidian response is the fact that Plantinga's proposed defeater for naturalism is (in the ways just noted) more like the Freudian and memory implant defeaters than it is like the defeater from evil.

The reason the Freudian defeater for theism and the artificial-memory-implant defeater for W are successful is that they cause the person holding the threatened belief to have doubts about that belief's source. I agree that if defeaters of this sort can succeed in causing such doubt, they will be effective even if one has nonpropositional evidence in support of the belief that the source in question is reliable. However, it's important to recognize that there is no reason to think that such defeaters will be successful if the person holding the threatened belief does *not* come to have doubts about the trustworthiness of its source. So the question we need to ask ourselves is whether belief in X and Y in cases that are like the first three *always requires* the person holding the threatened belief to have doubts about Z? I say it doesn't.

The natural response to learning X and Y in the Freudian and memory implant cases is for the person who was holding B to have grave doubts about Z and, on the basis of such doubts, to retract B. Let's call this sort of response 'response 1'. The alternative response that I want to consider—response 2—is this: the person who was holding B accepts X and Y, agrees that they show that Z is unlikely, but continues to hold both Z and B, pointing out that X and Y have simply shown her that something she knows to be true—i.e., Z—happens to be unlikely on other propositions she believes. This person has no doubts about Z even though she accepts X and Y. If there are cases where this second sort of response is rational, then there are cases where a potential defeater that has the same structure as the Freudian and memory implant defeaters doesn't work. And if Plantinga's defeater for naturalism is one of those cases, then my Reidian response is successful after all.

But isn't it obvious that there *are* cases in which someone can believe propositions resembling X and Y without having doubts about a corresponding proposition resembling Z? If a person comes to learn that some proposition she thinks is true is unlikely on other things she believes, that doesn't au-

[26] Note, however, that the three *undercutting* defeaters (for G, for W, and for N) involve *rebutting* defeaters for the belief that the source of the belief that G or W or N is reliable.

tomatically require her to reject it. For example, consider X, Y, and Z in the straw-drawing case:

Straw-Drawing Case
X: The straw is drawn from the lottery straws.
Y: P(straw is about three inches long/it is drawn from the lottery straws) is low or inscrutable.
Z: The straw that is drawn is about three inches long.

Suppose you came to believe, in the normal way (through observation), that Z—i.e., that the straw you just selected from a large number of lottery straws is about three inches long. Then suppose you learn that all the lottery straws but one were twelve inches in length and that, as a result, you believe Y. This together with your belief that X would *not* make it reasonable for you to stop believing that the straw you drew was about three inches long. Instead, it would make it reasonable for you to think that something you *know* is true (i.e., Z) is *unlikely* given your propositional evidence. This shows that there are cases where someone could believe X and Y without coming to have doubts about Z. And that suggests that there are cases where response 2 is the correct response to defeaters with the same structure as the Freudian and memory implant defeaters.[27]

But is Plantinga's proposed defeater for naturalism a case of this sort? Is it more like the straw-drawing case or more like the cases of the Freudian and memory implant defeaters? Well, in each of these four cases (the straw-drawing case, the Freudian and memory implant defeaters, and Plantinga's proposed defeater for naturalism) we have some evidence *against* Z. That evidence is a probabilistic inference from X and Y. If that evidence is not outweighed by other sufficiently strong evidence *for* Z, then the sensible thing to do is to have doubts about Z. The reason it is sensible to persist in believing Z in the straw-drawing case is that the evidence for Z from sense perception and memory is sufficiently strong to outweigh the counterevidence from the probabilistic inference. But in the Freudian and memory implant cases, there is no evidence for Z that outweighs the probabilistic evidence against it. In the remainder of this section I will explain why, in light of the Reidian account of how we know R, Plantinga's proposed defeater for

[27] One might resist the parallel I'm drawing between the straw-drawing case and cases like the Freudian and memory implant defeaters on the grounds that the latter two involve undercutting defeat whereas the straw-drawing case involves only rebutting defeat. But (as I mentioned in note 26) the cases involving undercutting defeat work by way of a rebutting defeater for Z—the claim (in the Freudian and memory implant cases) that *the source of the belief threatened with undercutting defeat is reliable*. And my point is that if this rebutting defeater doesn't work, then the undercutting defeater that depends on its success won't work either. So it *is* relevant to focus on rebutting defeaters (like the straw-drawing case) that parallel the sort of rebutting defeater that plays a role in the undercutting defeat in the Freudian and memory implant cases.

naturalism is more like the straw-drawing case than it is like the Freudian or memory implant cases.

Let's first compare the various cases with respect to evidence *against* Z. The natural way to understand the case of the Freudian defeater for theism is so that the person holding the theistic belief thinks that the probability mentioned in Y (i.e., the probability that a source is reliable given that it is a sort of wish fulfillment) is equal to zero. This makes one wonder whether the propositional evidence against Z in this case is properly thought of as merely *probabilistic*. For the person holding the threatened theistic belief doesn't think only that beliefs formed by way of wish fulfillment are *unlikely* to be reliable. She thinks they're *not* reliable. And, of course, if she thinks that then she should certainly have doubts about Z. Similar remarks apply to the memory implant case. If a person comes to believe that her memories are implants, then she just assumes their source isn't reliable.[28] But with Plantinga's proposed defeater for naturalism the case against Z isn't nearly as strong. In order for it to be as strong, Plantinga would need to argue that cognitive faculties that come about by way of N&E are *not* reliable, not merely that they are *unlikely* to be reliable (or that the probability that such faculties are reliable is low or inscrutable). But Plantinga doesn't come close to arguing that such faculties aren't reliable. In fact, the naturalist isn't even given a good reason to believe that the probability mentioned in Y in the Plantingian N&E case is *low*. The only persuasive reason Plantinga gives for thinking that P(R/N&E) is low or inscrutable is that P(R/N&E) is inscrutable.[29] This isn't to say that Plantinga's proposed defeater doesn't count against naturalism. If one was considering whether or not to maintain her belief in naturalism and the *only* relevant consideration was that the probability of that belief's source being reliable was inscrutable, that would be enough to make a sensible person have doubts about the source and, therefore, about naturalism. Nevertheless, it isn't a very *strong* reason to have doubts about Z—not nearly as strong as the reasons for doubts about Z in the Freudian and memory implant cases.[30] In fact, the Plantingian N&E case against *his* version of Z is weaker even than the (easily overturned) case against the Z of the straw-drawing example since, as I described in the latter example, the probability in question was known to be low.

[28] My point is that our *natural tendency* is to think that the subjects in these two cases who think their belief's source is wish fulfillment or due to a memory implant will think the source in question is *not* reliable, not merely that it is unlikely to be reliable. This tendency on our part (to read these particular cases non-probabilistically) is present even when the case is explicitly described in merely probabilistic terms.

[29] Plantinga suggests this in *WPF*, 229.

[30] This claim—that the evidence against Z in the Plantingian N&E case is weaker than the evidence against Z in the Freudian and memory implant cases—does *not* depend on my taking the probabilities mentioned in Y in the latter two cases to be equal to zero. So long as the probabilities in the Freudian and the memory implant cases are sufficiently low and the probability in Plantinga's N&E case is merely inscrutable, the claim is justified.

Now what about the evidence *for* Z? In the straw-drawing case, the evidence for Z is nonpropositional evidence via sense perception and memory. This evidence is strong enough to adequately counter the probabilistic inference against Z. It's not that the evidence for Z is a defeater-defeater that defeats the defeater constituted by the probabilistic inference from X and Y. Instead, the nonpropositional evidence for Z in the straw-drawing case prevents the propositional evidence against Z from ever functioning successfully as a defeater in the first place. The same sort of thing happens in the case of Plantinga's proposed defeater for naturalism. In the Plantingian N&E case, the naturalist has extremely strong nonpropositional evidence for Z—stronger, one might think, than the evidence one has for Z in the straw-drawing case. Z, in the Plantingian N&E case, is R; and the nonpropositional evidence supporting it comes by way of common sense. Again, it's not that the nonpropositional evidence from common sense is a defeater-defeater that defeats the probabilistic defeater for Z. Rather, the nonpropositional evidence for Z (i.e., for R) prevents the probabilistic inference from ever functioning as a defeater for Z in the first place.[31] Furthermore (and this is the important thing to recognize here), the evidence for Z in the Plantingian N&E case is as strong as or stronger than it is in the Freudian and memory implant cases. For it is natural and sensible to place much more confidence in R than in the reliability of the source of one's belief that G or the trustworthiness (on that particular occasion) of the source of one's belief that W. As a result, less counterevidence is required to make it rational to mistrust the source of G or of W than is required to make it rational to doubt R.

So, the evidence *against* Z is much weaker in the Plantingian N&E case than it is in either the Freudian or memory implant cases; it is even weaker than it is in the straw-drawing case. And the evidence *for* Z is much stronger in the Plantingian case than it is in either the Freudian or memory implant cases; it is at least as strong as it is in the straw-drawing case. It is, therefore, reasonable to conclude that in the Plantingian case the evidence for Z outweighs the evidence against it. In this way, the Plantingian case is more like the straw-drawing case than it is like the Freudian or memory implant cases. In the straw-drawing case, it is reasonable to believe that the straw you drew was about three inches long. All that your acceptance of X and Y does in that case is make it reasonable for you to think that something you know to be true—namely, that the straw you drew was about three inches long—is unlikely to be true given other things you know. The same goes for the case of Plantinga's proposed defeater for naturalism. It is reasonable for the naturalist to believe R. And all that her acceptance of X and Y does in that case is

[31] If the probabilistic evidence against Z *did* function as a defeater, the nonpropositional evidence for Z could not be used as a defeater-defeater of that defeater. For once one doubts Z (which in this case is R), one has an undercutting defeater for *all* one's beliefs, including R itself. One couldn't very well rely on common sense to remove one's doubts about common sense. See section 6 for further discussion of this issue.

make it reasonable for her to think that something she knows to be true—namely, that her cognitive faculties are reliable—is unlikely to be true given other things she believes.

In sum, the rational response to learning X and Y in the case of Plantinga's proposed defeater for naturalism is not response 1—i.e., having doubts about Z—but response 2—i.e., continuing to believe Z (which is R) despite the fact that it is known to be unlikely given other things one believes. The linchpin of my defense of this Reidian response to Plantinga's proposed defeater for naturalism is that our confidence in R (which is believed in the basic way) is reasonably as great or even greater than our confidence in properly basic sense perceptual and memory beliefs like Z in the straw-drawing case. It is because of this that the naturalist's belief in R is not defeated by the sort of probabilistic inference on which Plantinga relies in his evolutionary argument against naturalism.

VI. Epistemic Circularity

One might object to the parallel I've drawn between the straw-drawing case and Plantinga's proposed defeater for naturalism on the grounds that, in the straw-drawing case, Z is not about the trustworthiness of a belief's source whereas it is in the Plantingian N&E case. But that difference is irrelevant. Consider the case of having one's auditory beliefs challenged where X, Y, and Z are as follows:

Auditory Belief Source Case
X: My auditory beliefs are the result of input by scientists who, as far as I know, have no particular interest in the beliefs so produced being true.
Y: P(a belief source is reliable/it consists of input by scientists who, as far as I know, have no particular interest in the beliefs so produced being true) is low or inscrutable.
Z: My auditory beliefs are reliable.

If I ground my belief in Z in this case on the positive results of the many tests I and others have performed to check my hearing (tests which in no way depend on the reliability of *my* auditory beliefs) then that belief is completely reasonable. And this is so even if I also justifiably believe X and Y. I am completely justified in thinking that Z in this case is true despite the fact that it is unlikely to be true given the above versions of X and Y. The results here are exactly parallel to those in the straw-drawing case. So the alleged problem with the parallel I've drawn doesn't seem to be a problem at all.

But perhaps the worry about the difference between the straw-drawing example and the case of Plantinga's proposed defeater for naturalism is not just that Z, in the Plantingian case, is about the trustworthiness of a belief source whereas it isn't in the straw-drawing case. Perhaps the real concern with my

claim that the two cases are alike is that there is some sort of circular reasoning going on in the Plantingian N&E case that isn't present in the straw-drawing example (or in the auditory belief source case). In the straw-drawing case, what makes the belief that Z (i.e., that you drew a three-inch straw) reasonable even after you learn that it is extremely unlikely for such a belief to be true is the nonpropositional evidence you have, via your short-term memory and your sense perception, of your clearly observing yourself drawing that three-inch straw a few moments ago. But in the case of Plantinga's proposed defeater for naturalism, what makes your belief in Z (i.e., in R) reasonable, according to Reid, is the nonpropositional evidence you have for it via common sense. The problem is that in learning X and Y in the Plantingian N&E case you thereby obtain a reason to think that the probability that your belief sources—including common sense—are reliable is low or inscrutable. And, say those who object to my Reidian response, once you learn *that* you can't continue to trust your nonpropositional evidence for Z. For that evidence will itself be in doubt. You can't use the evidence of common sense to verify that the evidence of common sense (along with other kinds of evidence) is worthy of your trust.

As I suggested above, the complaint being expressed here is that circularity of a certain sort is unacceptable. You can't come to believe that a belief source is trustworthy by relying on the trustworthiness of that very source. This complaint sounds very plausible. But it is ambiguous. I will distinguish three ways of understanding it. The first two are extremely sensible; the third isn't. But it is only the third that applies to my Reidian response to Plantinga's proposed defeater for naturalism.

The first way to understand the complaint is as follows: a *logically* circular argument for the conclusion that a belief source is trustworthy contributes nothing to the reasonableness of accepting that conclusion. A logically circular argument is one that includes its conclusion among its essential premises (where an essential premise is one whose deletion would make the argument invalid). We needn't pause to discuss the reasons why such an argument is completely unhelpful.

The second way to understand the complaint is this: *epistemically* circular arguments for the trustworthiness of a belief source cannot help us *restore lost confidence* in such a source. An epistemically circular argument for the trustworthiness of a belief source S relies on premises, belief in which is produced by S. But if one has doubts about the trustworthiness of S, one's confidence in S won't be restored by an argument depending on S's own testimony. On this second reading, the complaint is one we can, once again, readily appreciate.

But neither of these two understandings of the complaint applies to the Reidian response I've developed.[32] According to that Reidian response, a per-

[32] Reid himself demonstrates an appreciation of something like the above two versions of the complaint in question. And his response to them bears a strong resemblance to the one I present in this paragraph (or perhaps I should say that my response bears a strong resemblance to his).

son can rationally continue to believe Z—which, in the Plantingian N&E case, is R—even after learning that it is unlikely on other things she believes. The reason belief in R can be maintained is that it is justifiably believed quite firmly in the *basic* way. By employing her faculty of common sense, the person holding R comes to have very powerful *non*propositional evidence for R, which justifies her belief in it. So R isn't believed on the basis of a *logically* circular argument. It isn't held on the basis of an argument at all since it is based on nonpropositional evidence. For the same reason, it isn't based on an *epistemically* circular argument.

Nevertheless, the belief in question *is* infected with a sort of epistemic circularity. For the belief that R—i.e., that all of our belief sources, including common sense, are reliable—is itself produced in the basic way by common sense. But although this is a case of a belief source vouching for its own trustworthiness (as well as the trustworthiness of many other belief sources), it is not a case of a belief source being depended on to *restore lost confidence* in its own trustworthiness. For the person in question does not have any doubts about the trustworthiness of common sense. It's true that if she did, then common sense would be of no help to her. But because she holds Z with confidence, there is no restoration of confidence needed. The epistemic circularity infects the process by which she comes to have confidence in Z in the first place. But it does not infect any process by which she comes to have restored confidence in Z since there *isn't* any such process. There isn't any such process because confidence in Z has not been lost and so is not in need of restoring.

This leads us naturally to the third and strongest reading of the complaint: if there is *any* epistemic circularity involved in the production or sustenance of a belief, that belief is not justified. And this complaint *does* apply to the Reidian response I've presented. For, as I explained in the previous paragraph, according to that response (to Plantinga), one can come to have, via common sense, a noninferentially justified belief that common sense is trustworthy.[33] But this complaint is not nearly so plausible as the other two. I grant that, at first glance, even this third version of the complaint seems to have something going for it. But a careful consideration of exactly what's involved in such circularity and of what the alternatives to it are makes it doubtful that this sort of circularity is vicious. I don't have the space here to delve into this. So I'll refer the reader to articles by Alston, Bergmann, Sosa, and Van Cleve for criticisms of the assumption that such epistemic circularity is always objectionable.[34]

See his discussion of the seventh of the "first principles of contingent truths" in Reid, *Essays*, 630–633 (parts of which are included in Reid, *Inquiries*, 275–277).

[33] Reid more or less explicitly concedes this. Concerning R, he says: "How then come we to be assured of this fundamental truth on which all others rest? Perhaps evidence, as in many other respects it resembles light, so in this also, that as light, which is the discoverer of all visible objects, discovers itself at the same time: so evidence, which is the voucher for all truth, vouches for itself at the same time" (Reid, *Essays*, 632).

[34] William Alston, "Epistemic Circularity," in *Epistemic Justification: Essays in the Theory of Knowledge* (Ithaca, N.Y.: Cornell University Press, 1989), 319–349; Michael Bergmann, "Exter-

However, I will point out one troublesome result of rejecting this sort of circularity. If this third complaint is an appropriate one, it creates problems not only for my Reidian response but also for anyone who thinks belief in R is justified. For, obviously, a person cannot come to believe R without relying on her own cognitive faculties. But that means that *any* belief in R will be infected with epistemic circularity. Sosa makes the point like this. Suppose that T is our total way of forming beliefs. The belief that *T is reliable* will, of necessity, be formed using T itself; any argument for T's reliability will be circular insofar as it employs premises that receive what warrant they have by being outputs of T. He concludes that:

> a correct and full response to rational pressure for disclosure of what justifies one in upholding the premises must circle back down to the truth of the conclusion. *Necessarily* such an argument must be epistemically circular—that much seems clear enough. To rue that fact at this stage is hence like pining for a patron saint of modesty (who blesses all and only those who do not bless themselves), once we have seen that there could not possibly be such a saint.[35]

Not even God could form the belief that his ways of forming beliefs are trustworthy without relying on those ways of forming beliefs.[36]

In short, my response to the third version of the circularity complaint is this. First, it is unconvincing for reasons that others and I discuss at length elsewhere. However, if it were plausible, we could easily show that *everyone*—naturalist and theist alike—has a defeater for R. For if the third version of the circularity complaint were correct, then belief in R can't be justified. The conclusion of Plantinga's evolutionary argument against naturalism would thereby be vindicated but at the cost of becoming highly uninteresting in light of a simpler argument for a much more general conclusion—one that creates a problem for Plantinga, the naturalist, and every other possible cognizer as well.

Before moving on, let me return to my claim that the circularity involved

---

nalism and Skepticism," *Philosophical Review* 109, 2 (April 2000): 159–194; Ernest Sosa, "Philosophical Scepticism and Epistemic Circularity," *Proceedings of the Aristotelian Society*, supp. vol. 68 (1994): 263–290; James Van Cleve, "Foundationalism, Epistemic Principles, and the Cartesian Circle," *Philosophical Review* 88 (1979): 55–91; and idem, "Reliability, Justification, and the Problem of Induction," *Midwest Studies in Philosophy*, vol. 9, *Causation and Causal Theories*, ed. Peter French, Theodore Uehling, and Howard Wettstein (Minneapolis: University of Minnesota Press, 1984). The general idea behind the defense of this sort of epistemic circularity is this. A belief can be justified according to a principle even if the person doesn't have a justified belief in the truth of the principle. The principle could be an internalist one like *a belief is justified if it is clearly and distinctly perceived* or an externalist one like *a belief is justified if it is formed via a reliable process*. To avoid regress problems, some internalists and most, if not all, externalists are willing to deny that the subject must believe the principle itself in order for the target belief to be justified in accordance with the principle. But then one could use beliefs that are justified *according to* the principle as premises in an argument that is intended as a justification for the conclusion that the principle in question is true. This implies that epistemic circularity is not always objectionable.

[35] Sosa, "Philosophical Scepticism," 284.

[36] See Plantinga, "Respondeo," 342.

in the Reidian response to Plantinga's evolutionary argument against naturalism is *not* of the unacceptable second kind—the kind where epistemic circularity infects the attempt to *restore lost confidence* in R. For someone might say this: "Once I come to believe that R is unlikely on other things I believe, I *should* lose confidence in my belief that R (in which case that confidence will need restoring after all). This is just to emphasize something that we accept in the straw-drawing case and the auditory belief case. In each of those cases, X and Y give us a reason, at least initially, to have doubts about Z.[37] But those doubts about Z are quickly countered (in those two cases) by strong evidence in support of Z—evidence whose value has *not been called into question*. The same thing should happen in the case of Plantinga's defeater for naturalism. If I am a naturalist, I should, initially at least, have doubts about Z (i.e., R) upon learning Plantinga's versions of X and Y. But, I can't in this case rely on any other evidence to counter these doubts because *all* other evidence has been called into question in virtue of my having doubts about R." The main point of this objection is that the believer in N&E *should initially have doubts* about R upon learning that P(R/N&E) is inscrutable. Because of these doubts, the epistemic circularity in question is of the unacceptable second kind since it is involved in the attempt to restore lost confidence in R.

So the crucial question is: Should the believer in N&E, initially at least, doubt R once she learns that P(R/N&E) is inscrutable? Or should her natural and sensible confidence in R remain high? If the latter, then the epistemic circularity is of the third kind and is not problematic; if the former, she is in real trouble.[38] To address these questions, let's consider another example:

Typical Skeptical Objection to R
Because of the following beliefs:
X: I currently have these experiences and beliefs
Y: P(R/I currently have these experiences and beliefs) is inscrutable
it is reasonable to doubt:
Z: R.

A few clarifying comments on this example are in order. First, the phrase "these experiences and beliefs" which appears in X and Y is to be understood

[37] Without pausing to pursue the matter, I'll just note that I disagree with this claim. I don't think it is true of either the straw-drawing case or the auditory belief case (assuming that in those cases we have the relevant evidence for Z *before* we learn X and Y).

[38] Notice that what's at issue here is whether the Reidian response to Plantinga's evolutionary argument against naturalism involves a defeater-defeater (this would be the case if Plantinga's argument led to initial doubts about R) or whether it prevents Plantinga's argument from functioning as a defeater for R in the first place (this would be the case if Plantinga's argument engendered no doubts). I agree that the Reidian response won't work as a defeater-defeater. If Plantinga's argument manages to create doubts about R, one can't rely on anything to quell those doubts (since all belief sources are now in doubt). But the Reidian response is not meant to be a defeater-defeater. As I argue below, it is intended to show that Plantinga's argument doesn't function as a defeater in the first place.

as if it is accompanied by an introspective pointing to *all* my current experiences and beliefs. Second, the motivation for believing Y in this example is due to reflection on typical skeptical scenarios like the Cartesian demon or the brain in the vat (hence the name of the example). The experiences and beliefs of a person who was the victim of a Cartesian demon or who was a brain in a vat would, by stipulation, be exactly as mine in fact are except that that person's belief sources would be radically unreliable. So how likely is it that one's belief sources are reliable given that she has the experiences and beliefs she does—ones that suggest to her (among other things) that her belief sources are reliable? It seems that, in light of the above considerations concerning typical skeptical objections, this probability is inscrutable. Hence Y above. (Of course if we added all the other things we believe to that on which R is conditionalized in Y, things might be different. But we aren't adding them.) Third, I agree that the following strategy of objecting to Plantinga's N&E argument isn't generally acceptable: (a) replace N&E in Plantinga's argument with *any* old proposition to get a parallel argument and (b) note that the parallel argument is a bad one. For the parallel argument to be acceptable for use in this sort of objection to Plantinga's argument, one must replace N&E with something that is *relevant* to our confidence in R. N&E is relevant because it tells of the origin of the faculties mentioned in R. But my replacement for N&E in the typical skeptical objection example—namely, that I currently have these experiences and beliefs—is also relevant because whatever it is that I have going for my belief that R is going to be among my beliefs and experiences (or at least many people think that this is so).

The thing to notice is that the typical skeptical objection to R given above doesn't require one to doubt R. One can recognize that the probability mentioned in Y (in the typical skeptical objection case) is inscrutable—that one can't see that it is high—and yet still sensibly believe that one's faculties are reliable. For one's belief that those faculties are reliable is, as Reid says, a *basic* belief. It is not formed on the basis of other beliefs such as the belief that *I currently have these beliefs and experiences* and that *R is likely given that I have these experiences*. Instead, belief in R is formed on the basis of the experiences themselves—in particular, the experience of the emotion of ridicule. (Sense perceptual belief works the same sort of way: I don't believe in the existence of external objects of sense perception on the basis of the belief that *I have such and such sense perceptual experiences* and that *the existence of such objects is likely given that I have experiences of that sort*. I just believe in such objects on the basis of the experience itself.) Because I don't base my belief in R on the belief that P(R/I currently have these experiences and beliefs) is high, the fact that I believe Y, according to which that probability is inscrutable, needn't lead me to have doubts about R.

In the previous paragraph, I've explained why, in the typical skeptical objection example, one can maintain belief in R despite one's acceptance of X and Y in that case. There is no need to doubt R, even initially. And my sug-

gestion is that since Plantinga's N&E case is relevantly like the typical skeptical objection example, the naturalist can sensibly maintain *without doubts* her belief in R despite her acceptance of Plantinga's versions of X and Y (having to do with N&E). Thus, the answer to our crucial question—Should the believer in N&E, initially at least, doubt R once she learns that P(R/N&E) is inscrutable?—is no. Just as rationality doesn't require us to doubt R in response to the typical skeptical objection given above, so also, the naturalist doesn't need to doubt R in response to Plantinga's evolutionary argument against naturalism.

If one wants to resist the response I've just recommended to the typical skeptical objection, then one can save Plantinga's argument but only at the cost of making it uninteresting. For using the typical skeptical objection, one can reach a similar skeptical conclusion (i.e., that R is defeated) that will apply to *everyone* regardless of whether he or she is a theist or a naturalist. In sum, either (a) there is no need to doubt R (even initially) in response to either Plantinga's N&E argument or the typical skeptical objection, in which case the only epistemic circularity involved in the Reidian response is of the acceptable *third* kind, or (b) one should doubt R in response to both Plantinga's N&E argument and the typical skeptical objection, in which case the epistemic circularity involved is of the unacceptable *second* kind. But if (b), then Plantinga's argument becomes uninteresting in light of the equally convincing typical skeptical objection for the same conclusion which applies to everyone—naturalist and theist alike.

VII. Naturalism and Common Sense

Let's turn now to a concern that a naturalist might have with the Reidian account of how we know R. The problem with Reid's account, according to this objection, is that it's not easy to see how to fit the faculty of common sense within a naturalistic framework. We can easily fit our sense perceptual faculties into this framework because we can see (at least in broad outline) how, in the usual cases, the fact that p causes our sense perceptual belief that p. And we can see that the relevant causal story involves no reference to things that aren't a part of the naturalist's ontology. But there is no available story of this sort to be told in the case of commonsense belief in propositions such as R. So my Reidian response to Plantinga isn't of much help to the naturalist.

A very similar complaint has been lodged against a priori knowledge.[39] But the sensible naturalistic response in each case—the first volley at any rate—seems to be essentially the same. First, we recognize that we do in fact seem

[39] Paul Benaceraff, "Mathematical Truth," *Journal of Philosophy* 70 (1973): 661–679; Devitt, "Naturalism and the A Priori," 45–65.

to know *in the basic way* (i.e., noninferentially) the very things that are supposed to be outputs of these faculties: I know in the basic way the necessary truths that modus ponens is a valid argument form and that 2 + 3 = 5; likewise, I know in the basic way the contingent truths that the thoughts of which I am conscious are *my* thoughts and that my cognitive faculties are reliable. But then, given this starting point, the naturalist will sensibly assume that, *unless there is some good reason to think otherwise*, common sense and a priori knowledge *are* compatible with naturalism.

Well, is there any good reason to think otherwise? Before answering this question, we need to highlight some of the many ways in which the term 'naturalism' is used. The reader will recall that, in this chapter, we have been concerned with *metaphysical* naturalism—roughly, the view that there are no supernatural entities.[40] One might well ask here whether the ontology of a metaphysical naturalist could include abstract objects like propositions, properties, numbers, or sets. Or should we just say that metaphysical naturalism is the view that everything is physical and, hence, that there are no abstract objects? Or perhaps there are several varieties of metaphysical naturalism—some of which include abstract objects in their ontologies and some of which do not. In any case, it is a sort of *metaphysical* naturalism that is the focus of our discussion in this chapter.

The other main sort of naturalism is *epistemological* naturalism. Alvin Goldman identifies three sorts of epistemological naturalism—scientistic naturalism (SN), empiricist naturalism (EN), and moderate naturalism (MN). He defines them as follows:

> (SN) Epistemology is a branch of science. The statements of epistemology are a subset of the statements of science, and the proper method of doing epistemology is the empirical method of science.
> (EN) All justification arises from empirical methods. The task of epistemology is to articulate and defend these methods in further detail.
> (MN)(A) All epistemic warrant or justification is a function of psychological (perhaps computational) processes that produce or preserve belief.
> (B) The epistemological enterprise needs appropriate help from science, especially the science of the mind.[41]

Plantinga's evolutionary argument against naturalism is not directed at these forms of naturalism. It is directed at a form of metaphysical naturalism. Like-

[40] One might object that this view is better called 'anti-supernaturalism' than 'naturalism' of some kind (I suspect Rea would say something like this—see *World without Design*). Perhaps that's true. But Plantinga ("Respondeo," 350–352) thinks that any sort of metaphysical naturalism will entail anti-supernaturalism. So in objecting to anti-supernaturalism, he is objecting to metaphysical naturalism too. In any case, whatever the view is called, it is the one Plantinga is attacking and I am defending.

[41] Alvin Goldman, "A Priori Warrant and Naturalistic Epistemology," *Philosophical Perspectives*, vol. 13, *Epistemology*, ed. James Tomberlin (Cambridge, Mass.: Blackwell Publishers, 1999), 2–3.

wise, my Reidian response is aimed at defending *metaphysical* naturalism against Plantinga's evolutionary argument. So what's important for my response is that knowledge via common sense is compatible with metaphysical naturalism, not that it is compatible with one of the above forms of epistemological naturalism. (Of course, metaphysical naturalism isn't compatible with Reid's ontology since Reid is a theist. But what's at issue is whether it is compatible with those elements of Reid's view on which my Reidian response to Plantinga's evolutionary argument depends.)

Now we can return to our question: Is there any good reason to think that a priori knowledge or commonsense knowledge is incompatible with naturalism? Prior to my laying out the above distinctions, one might have been inclined to say that there were some arguments on offer for the incompatibility of a priori knowledge and naturalism. But once we are sensitive to the fact that there are many different kinds of naturalism, we should be more hesitant in answering this question affirmatively. Michael Devitt and Philip Kitcher, two philosophers who take their naturalism to be in conflict with a priori knowledge, define naturalism so that is something like (E.N).[42] Devitt says that the heart of epistemological naturalism is captured by his claim that "the only way of knowing is the empirical way of knowing."[43] And Kitcher thinks of naturalism as including the claim that "virtually nothing is knowable *a priori*, and, in particular, no epistemological principle is knowable *a priori*."[44] These quotations certainly make it look as if they think of naturalism as the sort of epistemological naturalism that implies empiricism (or something very much like it). Of course, it is very easy to argue that *that* sort of naturalism is incompatible with a priori knowledge. But that isn't what is at issue in this chapter. Our concern is the compatibility of metaphysical naturalism with a priori knowledge and, more importantly, with commonsense knowledge. In light of the above comments, we shouldn't be so optimistic about there being good reasons for the naturalist to reject her initial assumption that metaphysical naturalism is compatible with these two kinds of knowledge. Even Devitt himself says:

> Metaphysical naturalism is *physicalism*: the view, roughly, that all entities are physical entities and that the laws they obey are in some way dependent on physical laws. This is a reductive doctrine. It has *nothing* to say about ways of knowing except that they must be, like everything else, physicalistically acceptable: so it alone entails nothing one way or the other about a priori knowledge.[45]

[42] Goldman makes basically this same point about Devitt and Kitcher. Ibid., 4.

[43] Devitt, "Naturalism and the A Priori," 46.

[44] Philip Kitcher, "The Naturalists Return," *Philosophical Review* 101 (1992): 53–114.

[45] Devitt, "Naturalism and the A Priori," 46. See also Georges Rey's defense of the compatibility of naturalism and a priori knowledge in "A Naturalistic A Priori," *Philosophical Studies* 92 (1998): 25–43.

The above remarks place the ball back in the objector's court. She could respond in a couple of ways. I will mention only two. First, she could disapprove of my taking it to be unproblematic to begin our epistemological investigations by appealing to obvious facts about what we know. In doing this I am employing what Chisholm calls 'the particularist approach' to epistemological questions.[46] Those who employ this approach begin with clear cases of what counts as knowledge or justified belief and proceed from there. But as I mentioned in the introduction, there are those who object to this approach.[47] Any naturalists who object to it will not be able to respond to Plantinga's arguments in the Reidian way I suggest. However, given the popularity of particularism among contemporary epistemologists—including contemporary epistemologists who are metaphysical naturalists—the Reidian response I've provided should recommend itself to many.

The other response the objector could make is to say that although metaphysical naturalism hasn't been shown to be inconsistent with commonsense knowledge, that doesn't give common sense a clean bill of health from the naturalist's perspective. For metaphysical naturalists are typically inclined to accept some sort of epistemological naturalism as well. So, according to the objector, I also need to show that commonsense knowledge is compatible with at least *some* sort of epistemological naturalism.

I'm not so sure this is true. But I offer this response. (MN)—the sort of epistemological naturalism Goldman favors—says the epistemological enterprise needs *appropriate* help from the sciences. Does this mean that, according to (MN), one can't posit *anything* in one's epistemological theory that hasn't been established by the sciences? This certainly doesn't seem to be the suggestion Goldman is making in proposing (MN). He insists that (MN) involves no such empiricist commitments.[48] John Pollock and Joseph Cruz also consider themselves to be supporters of epistemological naturalism. But they say this:

> At this stage of the development of cognitive psychology, the most appropriate methodology for investigating procedural knowledge is the standard philosophical methodology. The philosopher constructs thought experiments, makes judgments about whether certain kinds of beliefs or inferences would be epistemically justified in the circumstances envisioned, and then tries to construct a theory capturing the resulting "philosophical intuitions". . . . [T]he best way to

[46] Chisholm, "The Problem of the Criterion."

[47] Devitt says: "Epistemological naturalism [i.e., empirical naturalism or (EN) according to which the only way of knowing is the empirical way] applies to all knowledge. So it applies to knowledge of ways of knowing themselves" ("Naturalism and the A Priori," 46). This might signal his rejection of particularism insofar as particularism takes seriously (as a source of apparently nonempirical knowledge about our ways of knowing) our natural inclinations to judge certain cases of belief as obvious cases of knowledge.

[48] Goldman, "A Priori Warrant and Naturalistic Epistemology," 4.

> study [epistemology] may not be with the standard tools of contemporary cognitive psychology.[49]

Clearly, someone who is an epistemological naturalist in the sense that either Pollock, Cruz, or Goldman is could affirm the existence of a priori knowledge even if she lacked a scientific account of how a priori knowledge works.[50]

Now as I suggested above, one of the main reasons that some epistemological naturalists resist the suggestion that we have a priori knowledge is that we can give no plausible scientific account of it.[51] From the perspective of (EN) or (SN), this has devastating consequences for a priori knowledge (and a similar problem would afflict commonsense knowledge). For if we can't give a scientific account of a priori knowledge then, according to (SN), we can't do epistemology with respect to a priori knowledge. According to (SN), science is *the* method for doing epistemology. Likewise, if we haven't determined, using empirical methods (like those of science), that we have a priori knowledge then, according to (EN), we have no reason to think there is such a thing. For according to (EN), the *only* way to find things out is the empirical way. But from the perspective of Goldman's (MN) or Pollock's and Cruz's epistemological naturalism, the fact that we lack an adequate scientific account of a priori knowledge or of commonsense knowledge isn't as problematic. Sure, it would be great to have such accounts and that is where continued scientific investigation could help us. But although we don't yet have such accounts, we certainly haven't come close to exhaustively researching the phenomena associated with what we think of as a priori and commonsense belief. One can reasonably be hopeful that we will eventually come up with such accounts after further investigation. Indeed, as Goldman reports, there have recently been some advances in our scientific understanding of a priori knowledge.[52] Ultimately, however, nonempiricist and nonscientistic epistemological naturalists don't *need* such scientific results in order to take commonsense and a priori knowledge seriously.[53]

[49] John Pollock and Joseph Cruz, *Contemporary Theories of Knowledge*, 2d ed. (Lanham, Md.: Rowman and Littlefield, 1999), 172.

[50] One of the purposes of Goldman's "A Priori Warrant and Naturalistic Epistemology" is to argue that (MN) is compatible with a priori knowledge or warrant.

[51] Devitt, "Naturalism and the A Priori," 50.

[52] Goldman, "A Priori Warrant and Naturalistic Epistemology," sec. 5.

[53] Of course, those relying on common sense or a priori intuition *do* need to address potential threats to such knowledge if any are posed by such results (see the preface to *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, ed. Michael DePaul and William Ramsey [Lanham, Md.: Rowman and Littlefield, 1998]). But the threats have to be something more than evidence that a priori intuition or common sense is *fallible*. We know *that* of vision and memory and still sensibly rely on them. In considering the appropriate attitude to take toward common sense and a priori intuition, we don't want to follow the example of Descartes who, upon noting that his senses sometimes deceived him, made the effort to place no trust in them until their reliability was independently confirmed. Furthermore, the scientific results we look at have to be relevant to the trustworthiness of *common sense or a priori intuition*. It won't do to cite in this connection—as Edward Wisniewski does in "The Psychology of Intuition" (in *Rethinking*

Much more could and should be said about the compatibility of naturalism and Reid's view that we know things via common sense. But what I've said above shows, at the very least, that it is by no means obvious that the metaphysical naturalist (i.e., the anti-supernaturalist) cannot avail herself of my proposed Reidian response to Plantinga's evolutionary argument against naturalism.

VIII. But Is Reid's Account True?

Perhaps if Reid's account of how we know things via common sense were true it would make possible the response to Plantinga's argument sketched in section 4 and defended in sections 5 and 6. And perhaps, as I've argued in the previous section, Reid's account is *compatible* with metaphysical naturalism. But how does this help the naturalist if Reid's account isn't true or even plausible? In order for my commonsense response to Plantinga to be useful, we need some reason to take Reid's account seriously. I don't have the space here to launch into a full-fledged defense of Reid's account of commonsense knowledge. However, I would at least like to say *something* in support of Reid's view and to explain some of the considerations that attract me to it.

But first, let's make the task more manageable. There is no need to defend all of Reid's views on common sense. One certainly doesn't need to agree with Reid about which propositions are first principles. For the purposes of this chapter, what matters is that we justifiably believe R via common sense. Nor does one need to hold Reid's views on the details of how one comes to believe in first principles like R—details such as whether we have a faculty of common sense or whether there is an emotion of ridicule. What matters for my Reidian response to Plantinga is that *we believe R noninferentially on the basis of some sort of nonpropositional evidence and thereby have a lot of justification or warrant for it*. Our question, then, is whether this last (italicized) suggestion can be taken seriously.

As a matter of fact, it *is* taken seriously. I've mentioned above that the particularist approach to epistemology is currently quite popular among analytic philosophers. Those who employ it rely heavily on the noninferential knowledge they have of Moorean truths—truths such as that we aren't being de-

---

Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry, ed. Michael DePaul and William Ramsey [Lanham, Md.: Rowman and Littlefield, 1998], 51–57)—evidence of the fallibility of (a) philosophical theorizing about causes of language use or (b) introspective identification of the causes of our beliefs or (c) distinguishing on reflection our acts of memory from our acts of imagination. (Wisniewski seems to classify each of these activities as intuiting of some sort. But none of them fall into the category of things known via Reidian common sense or a priori intuition.) The problem is that the fallibility of the ways of forming beliefs mentioned in (a)–(c) isn't a good basis for concluding that common sense and a priori intuition are fallible—much'less for concluding that they are unreliable.

ceived by a Cartesian demon about the external world or about the past and that we aren't brains in vats. According to the particularist methodology, accounts of justification according to which it turns out that we aren't justified in believing such Moorean truths are, thereby, disqualified. The question of *how* it is that such epistemologists know these Moorean truths isn't very often addressed. But it seems that any answer given will be something along the lines suggested by Reid's account sketched above in section 3. At the very least, it will involve justified or warranted noninferential belief in R.

Furthermore, a defense of Reid's account could be developed along the following lines. First one could point out that sense perceptual beliefs based on sense experiences can be justified despite the fact that we lack compelling noncircular inductive or deductive arguments from the existence of such experiences to the truth of the beliefs they ground.[54] Then one could point out that a priori knowledge also involves belief processes in which a belief is based on a certain sort of seeming—a seeming which is an experience of some sort.[55] In this case too it looks like there is no deductive or inductive argument from the existence of such an experiential ground to the truth of the a priori belief based on it. But this doesn't cast doubt on the justification of our a priori beliefs any more than a similar concern casts doubt on the justification of our sense perceptual beliefs. In each case (sense perceptual and a priori) the belief is noninferentially justified as a result of its being based on the experiential ground in question.

Once one has shown that the above suggestions are plausible one could then argue that commonsense belief in contingent truths is very much like a priori belief insofar as they each have the same sort of experiential ground (i.e., a certain sort of seeming). Because it is plausible to take seriously both the existence of justified a priori beliefs as well as the account of them as experience based, it is also plausible to take seriously both the existence of justified commonsense beliefs in contingent truths as well as an account of them as experience based. Our starting point is the fact that we *do* seem to have justified noninferential beliefs of each kind despite the fact that in each case the belief in question is based on a ground from whose existence we can't deductively or inductively infer the truth of the belief it grounds. And my suggestion is that since many philosophers are inclined to express very little resistance to an account of justified sense perceptual belief according to which the justifying grounds don't entail the truth of the belief, there should also be very little resistance to similar accounts of justified a priori or commonsense belief—especially when such accounts fit so nicely with our introspective un-

[54] See William Alston, *The Reliability of Sense Perception* (Ithaca, N.Y.: Cornell University Press, 1993) for an extended critique of various attempts to show that there is some such connection between sense experiences and the beliefs they ground.

[55] See Plantinga, *WPF*, chap. 6. Because the experiential ground of a priori beliefs is nonempirical the beliefs are still properly called 'a priori' rather than 'a posteriori'.

derstanding of what is going on in typical cases of what seem like justified a priori or commonsense beliefs.

The above remarks are meant to gesture in the direction of a defense of that part of Reid's account that is employed in my response to Plantinga. They are *not* intended to put to rest all doubts those who resist Reid's account might have. To do that one would have to defend the existence of a priori belief as well as the account of it as experience based. And one would also need to consider whether the differences between commonsense belief and a priori belief prevent us from moving from a favorable evaluation of our account of justified a priori belief to a favorable evaluation of a similar account of commonsense belief. For example, one might think (though I don't) that although we *can* have noninferential knowledge of propositions that are general and necessary as well as of propositions that are particular and contingent, we *can't* have noninferential knowledge of propositions that are general and contingent. If this were true, it would suggest that a priori knowledge (which is typically of truths that are general and necessary) is acceptable in a way that commonsense knowledge (which at least sometimes is supposedly of general and contingent truths like R) is not. But dealing with these sorts of concerns is a project for another occasion. Here I merely hope to have shown that a plausible defense of the required elements of Reid's account of commonsense knowledge is by no means out of the question.

IX. Skepticism

In closing, I would like once again to draw attention to the fact that Plantinga's argument against naturalism can also be thought of as an argument from naturalistic premises for global skepticism. Consequently, in defending and developing my Reidian response to Plantinga's argument against naturalism, I have thereby been mustering support at a more general level for a commonsense response to skeptical challenges of various kinds. In particular, I've explained (a) the role that reliance on nonpropositional evidence plays in a commonsense account of how we know that our faculties are reliable (see sections 3–5), (b) the reason the commonsensist needn't be troubled by the charge of epistemic circularity (see section 6), (c) the reason the naturalist *qua* anti-supernaturalist can feel comfortable with this commonsense response to the skeptic (see section 7), and (d) the reason one can take seriously certain crucial elements of Reid's account of how we know R in the basic way (see section 8). All of these points are important in a Reidian response to skepticism. Of course, they could all use further development. And their application to the general issue of skepticism needs to be more explicitly spelled out. But that too is a project for another paper. Here I simply want to highlight the broader context in which my Reidian response to Plantinga should be sit-

uated; it is just one application of a commonsense response to skepticism. I recognize that within that broader context the commonsense response won't satisfy the skeptic. However, I (like Reid) am quite skeptical of the claim that a philosophically adequate and respectable response to the skeptic must leave her satisfied.

Plantinga's Evolutionary Meditations

Ernest Sosa

Humans can rise to a higher level of knowledge, reflective knowledge, governed by this Principle of the Criterion:

> PC Knowledge is enhanced through justified trust in the reliability of its sources.[1]

This is meant in the first instance as a principle about the contents of a single stretch of consciousness in one specious present when one actually considers the reliability of one's sources for a given belief. But it can also be extended to cover one's implicit beliefs, which rise to the higher level only by meeting the requirement that under the light of reflection one must be able justifiably to defend the adequate reliability of one's sources. Aside from unreflective, animal knowledge, therefore, humans aspire also to a higher knowledge, reflective knowledge, which obeys our principle of the criterion.

Our conception of reflective knowledge is interestingly and closely related to Alvin Plantinga's extended argument for the claim that naturalism is self-defeating, an argument that he has had under development for several years now, through multiple presentations and much controversy. Here I would like to examine this argument and compare it with our conception of reflective knowledge. Sections 1 and 2 will develop an account of such reflective

[1] I have defended such a principle in "How to Resolve the Pyrrhonian Problematic: A Lesson from Descartes," *Philosophical Studies* 85 (1997): 229–249. More fully stated, the principle holds the epistemic quality of one's belief to rise with justified awareness of the reliability of one's sources.

knowledge which I find strongly suggested by Descartes, and sections 3 and 4 will discuss Plantinga's argument in its light.[2]

I

Early in his Meditations, Descartes is struck by the troubling consequences of our "principle of the criterion," which may be put more simply as follows:

> PC High-level knowledge requires justifiedly taking one's sources to be reliable.[3]

Recall his desire to counteract doubt that one's operative sources or faculties may be unreliable—thus the doubts deriving from the fallibility of the senses, from the possibility that one is dreaming or has fallen victim to an evil demon, etc. The problem in each case is that the source of one's belief might be unreliable: that is, might not with sufficient reliability protect one from error. And note well Descartes's reaction: in order to remove such doubts, he wishes to establish that our beliefs do in fact derive from sources reliably worthy of our trust. This might be thought to derive only from his aim to prove the reliability of our sources, *without* implying that we attain true knowledge only once we achieve that aim. How that misinterprets Descartes's intentions comes out in the following passage:[4]

> The fact that an atheist can be "clearly aware that the three angles of a triangle are equal to two right angles" is something I do not dispute. But I maintain that this awareness of his [*cognitionem*] is not true knowledge [*scientia*], since no act of awareness that can be rendered doubtful seems fit to be called knowledge [*scientia*]. Now since we are supposing that this individual is an atheist, he cannot be certain that he is not being deceived on matters which seem to him to be very evident (as I fully explained). And although this doubt may not occur to him, it can

[2] The interest of thus comparing Plantinga's antinaturalist argument is likely to depend on the interest of our conception of reflective knowledge itself, which I will try to defend in detail with the help of some historical perspective.

[3] A principle best implicitly relativized to those who grasp (understand) the proposition that they know that p.

[4] This passage is important enough that I would like to have before us an alternative translation: "That an atheist can clearly know that the three angles of a triangle are equal to two right angles, I do not deny; I merely say that this knowledge of his (*cognitionem*) is not true science (*scientia*), because no knowledge which can be rendered doubtful should, it seems, be called science. Since he is supposed to be an atheist, he cannot be certain that he is not deceived even in those things that seem most evident to him, as has been sufficiently shown; and although this doubt may never occur to him, nevertheless it can occur to him, if he examines the question, or it may be suggested by someone else, and he will never be safe from it, unless he first acknowledges God." *The Philosophical Works of Descartes*, trans. Elizabeth S. Haldane and G. R. T. Ross (Cambridge: Cambridge University Press, 1911), 2:39. Cf. *Oeuvres de Descartes*, ed. Charles Adam and Paul Tannery (Paris: Librairie Philosophique J. Vrin, 1964), 9:111.

still crop up if someone else raises the point or if he looks into the matter himself. So he will never be free of this doubt until he acknowledges that God exists.[5]

Principle PC sufficiently explains why Descartes will not be satisfied with unqualified foundationalism. After all, enlightened, reflective *scientia* requires the satisfaction of PC, and will not be attained through mere external foundations in the dark, nor even when the foundational source is "internal" and as reliable as rational intuition is said to be.

Descartes will not settle for mere *cognitio*, not even for *internalist*, a priori, reason-derived cognitio, as attained by the atheist mathematician. Descartes wants *reflective*, *enlightened* scientia. It is *this* that sets up the problem of the Cartesian Circle. Since Descartes wants not just reliable, truth-conducive cognitio, since he wants the enlightened attainment of reflective scientia, he needs a defense against skeptical doubts that target his intellectual faculties, not only his faculties of perception, memory, and introspection, but even his faculty of intuitive reason, by which he might know that $3 + 2 = 5$, that if he thinks then he exists, and the like. He thinks he can defend against such doubts *only* by coherence-inducing theological reasoning yielding an epistemic perspective on himself and his world, through which he might confidently trust his faculties. And these faculties must include those employed in arriving, via a priori theological reasoning, at his perspective on himself and his world, the perspective that enables confidence in the reliability of those very faculties.[6]

We may explicate Descartes's project by placing it in the context of the Pyrrhonian problematic. This also helps explain why the circle is virtuous, and how certain stages of the Cartesian project, which might seem incoherent at first blush, are defensibly coherent in the end. (Example: the apparently incoherent claim about needing to first prove the veracity of God.)

In the barest sketch, here is how I see Descartes's epistemological project. First he meditates along, attaining the kind of epistemic justification and even "certainty" that might be found in an atheist mathematician's reasonings, one deprived of a world view within which the universe may be seen as epistemically propitious. Descartes's reasoning at that stage *can* be evaluated, of course, just as can that of the atheist mathematician. After all, atheist mathematicians will differ in the worth of their mathematical reasonings. Absent an appropriate world view, however, no such reasoning can rise above the level

[5] This passage is from the Second Set of Replies as it appears in *The Philosophical Writings of Descartes*, trans. J. Cottingham, R. Stoothoff, and D. Murdoch (New York: Cambridge University Press, 1984–91), 2:101. Where this translation says that an atheist can be "clearly aware," Descartes's Latin is *clare cognoscere*.

[6] There is a telling analogy, not explicitly recognized by Descartes, between the role of dreams in his skepticism vis-à-vis perception, and a role assignable to paradoxes and aporias in a parallel skepticism vis-à-vis rational intuition.

of cognitio. If we persist in such reasoning, nevertheless, enough pieces may eventually come together into a view of ourselves and our place in the universe, a view sufficiently comprehensive and coherent to raise us above the level of mere cognitio and into the realm of higher, reflective, enlightened knowledge, or scientia. There is in none of that any circle that vitiates the project.[7]

Admittedly I have done little more than to sketch my proposed interpretation of Descartes's epistemological project. A proper defense would require a full textual study. Some key texts inspire optimism, but the full project must await another occasion.

One important question I would like to mention before moving on: What could possibly give to reflective knowledge a higher epistemic status than the corresponding unreflective cognitio(n)? And, more particularly, what could possibly do so within the epistemological framework favored by Descartes?

What favors reflective over unreflective knowledge? Reflective acquisition of knowledge is, again, like attaining a prized objective guided by one's own intelligence, information, and deliberation; unreflective acquisition of knowledge is like lucking into some benefit in the dark. The first member of each pair is distinguished from the second in being more admirable, something that might be ascribed admiringly to the protagonist, as his doing. And we can after all shape our cognitive practices, individually and collectively, enhancing their epistemic virtue, their effectiveness in putting us in touch with how matters stand. We can do so at least to some extent, which does not require that our every belief be freely chosen and deliberate. A tennis champion's "instinctive" reactions at the net derive from a highly deliberate and autonomously chosen training carried out voluntarily over a period of years. Even when already in place, moreover, such "instinctive" reactions are still subject to fine tuning through further practice and training. The same is true of a bird watcher and his binocular-aided "instinctive" beliefs. And the same is true of us all and our most ordinary visual beliefs, aided by the tutelage of daily practice and, eventually, the hard lessons of diminishing acuity.[8]

[7] Among the pieces that need to come together in order to raise the belief-that-p above the level of cognitio, to the level of scientia, may well be found appropriate cognitio *that* one enjoys cognitio that p. And once certain claims of priority are dropped, as I am proposing, then it might well be held that cognitio that p *and* cognitio that one enjoys cognitio that p are both required for scientia that p. It might even be held that scientia that one has scientia that p is also required for scientia that p—so long as one grasps the proposition that one has scientia that p. So a form of the KK Principle seems accessible along this avenue.

[8] It must be granted, however, that Descartes's cognitio-level attainments are importantly different from the lucky grasp of gold by those in the dark room. The latter is more like a gambler's lucky guess. Undeniably, that distinction can and should be made. But it brings to the fore the fact that epistemic luck can be found at different levels and in different ways. Take a case where the gold searcher enters a dark room that in fact happens to contain only gold objects, though he has no idea of this. Or take a case where there are several rooms before him, any of which he might enter, all dark and only one containing gold objects. If he happens to choose that one while ignorant of the relevant facts, again it is no accident in one respect that he lays hold of some gold, but it remains an accident in another respect. The Pyrrhonian gold-in-the-dark example can, I

A further advantage of reflective knowledge is its entailed increment of comprehensive coherence, something accepted by Descartes himself as a source of epistemic worth, indeed as a source of certainty. In Principle 205 of his *Principles of Philosophy*, for example, he notes that if he can interpret a long stretch of otherwise indecipherable writing by supposing that it is written in "one-off language," with the alphabet all switched forward by one letter, etc., this supports that interpretation. There he defends his account of physical reality in that "it would hardly have been possible for so many items to fall into a coherent pattern if the original principles had been false."

Since the propositions that form the coherent pattern are "deduced" from the "original" principles, therefore if the coherence of the implicands signifies the truth of the principles, it likewise, derivatively, signifies the truth of the implicands themselves forming the pattern. Descartes seems therefore here again, as elsewhere, to be adopting a kind of reliabilism in defending the epistemological power of coherence—its ability to impart certainty—by appeal to its reliability as a source of the truth of what renders it certain.[9]

Admittedly, in that same Principle 205 Descartes claims only *moral* certainty for his coherence-validated beliefs about the alphabet or about the natural world. However, in the very next principle (206) he claims *more* than moral certainty for his scientific principles, and does so, again, at least in part through appeal to explanatory coherence (when he adduces that they "appear to be the only possible explanations of the phenomena they present").

II

Ancient skepticism, as represented by Pyrrhonism, and modern skepticism, as presented by Descartes, have been regarded as radically different. How plausibly? One problem raised by Descartes *is* limited by comparison with the radical skepticism of the ancients: namely, the problem of the external world.

---

believe, be supplemented interestingly in the ways indicated, so as to support the intuition that one needs a reflective perspective rich and powerful enough to rule out more than just the weak possibility that one's belief is just as much a matter of blind luck as is a gambler's lucky guess.

[9] This appeal is also found at the heart of Descartes's epistemology, when in the second paragraph of the Third Meditation he writes: "Certainly in this first knowledge there is nothing that assures me of its truth, excepting the clear and distinct perception of that which I state, which would not indeed suffice to assure me that what I say is true, if it could ever happen that a thing which I conceived so clearly and distinctly could be false." Reliability is hence assumed to be at least necessary in an acceptable source of epistemic status. A belief is supposed to be certain if one is assured of it by its clearness and distinctness, which requires that clearness and distinctness be a (perfectly) reliable guarantor of truth. Note well: it is the clearness and distinctness of the perception that itself yields the certainty, at least in the first instance. What yields the certainty is *not* just an argument that *attributes* in a premise clearness and distinctness to one's perception of what one is led to accept as a conclusion of that argument. (As we have seen, such an argument may *boost* the epistemic status of one's belief in its conclusion, by making it a case of reflective scientia. But that belief already enjoys the highest level of certainty attainable as a state of cognitio simply through the perception of what it accepts with sufficient clearness and distinctness.)

But this is not the only skeptical problem of interest to Descartes. It is obvious in the *Meditations* that his concerns are much broader, as when he wonders how he can know the truth even when he adds three and two or when he considers how many are the sides of a square. I have argued that it is precisely the radical skepticism of the ancients that mainly concerns Descartes (and not only Hegel, who is emphatic on the point). Moreover, this skepticism is best seen in the light of the epistemic problematic found already in Aristotle's *Posterior Analytics*,[10] where it is given a foundationalist resolution, and, more famously, in the five modes of Agrippa. To the latter incarnation of that problematic the Stoics, in kinship with Aristotle, offer a foundationalist response. Where Aristotle appeals to rational intuition as a way to found scientific knowledge, the Stoics appeal to natural, animal perception as a way to found ordinary empirical knowledge.

The Pyrrhonists reject such externalism because it dignifies mere "groping in the dark" with the title of knowledge. The Pyrrhonists highlight *enlightened* knowledge, which must be acquired and sustained in awareness of one's epistemic doings. Only this is "knowledge" worthy of the title. Sadly, they would prefer in their own practice to suspend judgment in specific case after specific case, partly because they reject blind foundations. In their view, moreover, any attempt to move beyond foundations only misleads us into circles or regresses, viciously either way.[11]

Descartes's response is balanced and sensitive to this (Pyrrhonian) dialectic. It grants the truth in foundationalism by allowing an inference-independent epistemic state of cognitio. Perception, for example, might well give us such unreflective animal knowledge unaided by inference.[12] Even intuition might give us foundational cognitio, as suggested by Aristotle, and such unreflective knowledge is open even to the atheist mathematician.

Such externalist epistemologies have prompted much controversy. Internalists have rejected them as unworthy of human epistemic dignity. Externalism has been denigrated as a "thermometer" model of knowledge inadequate to the full complexity of human cognition.[13] Despite recent decades of such

[10] A 3.

[11] The exceptions allowed—perhaps ordinary beliefs generally, as opposed to the theories of philosophers or scientists—would seem unmotivated if their best and deepest arguments would allow no such discrimination.

[12] My *Knowledge in Perspective* (Cambridge: Cambridge University Press, 1991) gives a central role to a distinction between animal knowledge and reflective knowledge analogous to Descartes's cognitio/scientia distinction, although my rationale is somewhat different and predates my awareness of Descartes's. See, for example, pp. 240 and 282. I would now supplement that earlier rationale with our considerations here involving reflective knowledge, the Principle of the Criterion and its demonstration, and the Pyrrhonian Problematic. These considerations support use of the animal/reflective distinction in epistemology, no less than they support Descartes's use of the cognitio/scientia distinction. Of course, many since Descartes have groped for a similarly coherent way beyond the Pyrrhonian Problematic, and many have repudiated any such project.

[13] Cf. Wilfrid Sellars, "Empiricism and the Philosophy of Mind," in *The Foundations of Science and the Concepts of Psychoanalysis*, Minnesota Studies in the Philosophy of Science, vol. 1, ed. H. Feigl and M. Scriven (Minneapolis: University of Minnesota Press, 1956), pt. 8, sec. 36 ("Does

controversy, we can still learn from our ancient predecessors, or so I will suggest. The Pyrrhonists oppose the externalism of Galen and the Stoics. Sextus, in particular, invokes similes that illuminate our issue, such as the following:

> Let us imagine that some people are looking for gold in a dark room full of treasures. . . . [N]one of them will be persuaded that he has hit upon the gold even if he *has* in fact hit upon it. In the same way, the crowd of philosophers has come into the world, as into a vast house, in search of truth. But it is reasonable that the man who grasps the truth should doubt whether he has been successful.[14]

Most would not disdain the good fortune of striking it rich in the dark, but it is no doubt a lesser state than that of finding gold through a deliberate plan aided by good eyesight in clear light. Enlightened discovery is more admirable than is any comparable luck that may reward groping in the dark. For one thing, enlightened discovery is success attributable to the agent; luck in the dark is not.[15]

Suppose that, concerning a certain subject matter, you ask yourself whether you know, and you have to answer either "Definitely not," or "Who knows?" In that case, is there not some straightforward and widely shared sense in which you do *not* really know?

Similarly, suppose that, concerning a certain choice or action of yours, you ask yourself whether it is right and you have to answer "Definitely not," or "Who knows?" Isn't there some sense in which your action or choice *automatically* falls short?

Such considerations may amount to nothing more than this: It is better to believe and to act in ways that are *reflectively* right than in ways that happen to be right but unreflectively so. There is a higher state of knowledge, reflective knowledge, but one subject to our Principle of the Criterion, PC above. Attaining such knowledge requires a view of ourselves—of our beliefs, our faculties, and our situation—in the light of which we can see the sources of our beliefs as reliable enough (and indeed as perfectly reliable if the scientia desired is absolute and perfect).

Why is such reflective scientia better than unreflective cognitio? First, because it is reason-molded, at least in the way of a champion's "instinctive" play. Second, because a knowledge that enjoys the support of a comprehensively coherent and explanatory world view is better. But is it *epistemically* better? Is it better with a view to getting at the truth? That Descartes would re-

---

Empirical Knowledge Have a Foundation?"), which I examine in "Mythology of the Given," *History of Philosophy Quarterly* 14 (1996).

[14] *Against the Mathematicians*, VII 52, in the Teubner text, ed. H. Mutschmann (Leipzig: n.p., 1914).

[15] In harmony with this Pyrrhonian sentiment, Confucius says in his *Analects*, bk. 2, no. 17: "To say you know when you know, and to say you do not when you do not, that is knowledge."

spond affirmatively even here is made plausible by key passages in which he explicitly recognizes the epistemic power of explanatory coherence.

In sum, Descartes was a foundationalist *and* a coherentist *and* a reliabilist.[16]

III

Naturalism eschews or rejects appeal to the supernatural and traces our origins back to blind and uncaring forces. Relative to such brute forces, however, the probability that our cognitive faculties are reliable must be either quite low or at best inscrutable. This defeats any belief we may have in the reliability of our faculties. Absent such belief, finally, we are deprived also of epistemic warrant (authority, justification) for all beliefs deriving from such faculties. But among these beliefs is the very belief in naturalism, which therefore defeats itself.

Reply: "Surely we naturalists know more about our origins than that! We know we derive from evolutionary processes that ensure our fitness to do well in our environmental niche." Unfortunately, this ensures at most that we excel in certain respects that help enhance the fitness and survival of our species, and it is unclear just how correct beliefs can bear on such success. Even given such evolutionary considerations, accordingly, the likelihood that our faculties are cognitively reliable is only low or inscrutable. Naturalism thus remains defeated and cannot be saved by evolution.

So argues the antinaturalist Plantinga, who now defends his argument against a broad and varied set of objections gathered around it in the years since it was first sketched in his *Warrant and Proper Function*.

For a defense beyond that initial sketch Plantinga now explores in some depth the epistemology of defeat. Candidates for defeat are beliefs, and candidates for defeaters are, in the main, other beliefs and experiences. Any belief B that is part of one's noetic structure N is defeated by a newly acquired belief D if, and only if, so long as one retains belief D one can preserve one's rationality only by giving up belief B. It would not be fully rational to retain the noetic structure that includes both the belief B and the belief D.

For example, if one somehow believes oneself at the epistemic mercy of a demon bent on mischief, this belief D_1 is a defeater for one's trust (R) that one's own faculties are reliable, and D_1 is also thereby a defeater in turn for the ostensible deliverances of these faculties. Similarly, argues the antinaturalist, if one comes to believe that one is a product of brute forces, this belief D_2 also defeats one's trust in one's faculties, and in turn defeats their ostensible deliverances. (Here I have simplified Plantinga's own actual argument by

[16] His form of reliabilism appears already in the *cogito* passage itself, in the Third Meditation, where he says that clarity and distinctness could hardly serve as a source of certainty if it could ever happen that something could be so clear and distinct while false.

dispensing with the trappings of evolutionary theory and going directly to the belief that brute forces account for our existence and epistemic constitution or character.)

How plausible is this?

Relative to their having derived from brute forces by evolutionary processes, it is at best inscrutable how reliable one's faculties are likely to be. Suppose that precludes our rationally believing our faculties to be reliable. So we are now granting for the sake of argument that our belief in the reliability of our faculties is indeed defeated by that fact about their derivation. Even so, why should that also defeat our many ordinary beliefs derived from those faculties. Take a child at a pre-reflective stage wherein one takes no notice of one's faculties. Is that child irrational in believing as it does about food, shelter, and other basic matters of its simple life? If not, then why should it be irrational to harbor such simple beliefs conjointly with a belief (call it DERIVATION) that we derive from brute forces relative to which the probability of our being reliable is low or inscrutable? Even granted we accept DERIVATION, and even granted this requires us to withhold on R, why should this lack of belief about one's faculties make it irrational to uphold our simple beliefs about food and shelter, if a similar lack does not make such beliefs irrational in a child?

It is here that Plantinga's antinaturalist argument may connect with the reflective knowledge tied to our Principle of the Criterion. It is not really irrationality that the child is doomed to by lacking a relevant epistemic perspective on its cognitive doings and their reliability. Rather is he denied any possibility to ascend from mere animal knowledge to the reflective knowledge that requires such perspective. Absent belief in the reliability of one's own faculties, one is denied reflective knowledge. But is the naturalist denied any possibility of achieving the epistemic perspective needed for reflective knowledge? Has the supernaturalist an advantage in this respect?

Of course the belief that one's faculties are quite *unreliable* would fit ill with accepting their ostensible deliverances. Here then is an implicit assumption perhaps used by Plantinga: that if our faculties derive from brute forces then it is quite *unlikely* that they are reliable. Absent some way of defeating this consideration, therefore, we must hold our faculties unreliable, and this would make it irrational to accept their deliverances at face value. This does require the stronger claim that our faculties are *unlikely* to be reliable if derived from brute forces (through evolution), and not just the claim that the likelihood of their reliability is inscrutable. But let us proceed on the basis of this stronger claim nevertheless, for the sake of argument.[17]

Suppose a walk down by the riverside turns up a smooth, round stone,

[17] Note how well the stronger claim comports with acceptance of the Argument from Design. But here we shall focus on *reflective knowledge* specifically. It is then plausible to require belief in the reliability of one's faculties as a condition for enjoying (reflective) knowledge through their ostensible deliverances. This is not to deny that there is also a kind of "animal" knowledge that

which one picks up and admires, and about which one notes that it would reliably roll down inclines. "That stone must have derived from brute forces," thinks one, however, as a good naturalist, "and relative to that fact there is only a low or inscrutable probability that it be either round or a reliable roller." Have we now in this consideration a defeater of the belief (call it S) that the stone one holds is indeed round and a reliable roller?

We can plainly see and feel the stone's smooth roundness, and we know through much experience that such an object would roll reliably. The improbability of its having been rounded so smoothly, given its origin in brute forces, is then, surely, no bar to our still knowing it to be smoothly round, and even a reliable roller, nor does our justification for so believing seem defeated even if we retain the belief that here the improbable has happened.

If that is a plausible response in the case of the round stone, why are we deprived of it when it comes to our own nature as reliable perceivers endowed with eyes and ears, etc., by means of which we have reliable access to the colors and shapes around us? Start first with someone else, a friend of whom we believe (F) that she is a reliable perceiver of the colors and shapes of things seen in good light, etc. Can't we forestall the alleged defeat by appeal to our perceptual and other means of knowing that our friend is as described?

Perhaps it's when one turns to one's own case that we run into real trouble?

Not if the belief concerns only some introspectible feature of one's present state of consciousness or some perceptible feature of one's body. I can know that I now ache or that I have hands, no matter how improbable it may be that brute forces should produce someone aching or with hands, and despite my believing us to derive from such forces.

So it must be something concerning the specific belief about oneself involved in R that might make one's belief in R susceptible to defeat in the ways noted by Plantinga. And what exactly could that be? What might be the basis for the conclusion or for the assumption that R specifically is thus susceptible? Here is a possibility.

One might suppose that a deliverance of a faculty when accepted is based on an implicit reliance on the reliability of the faculty, and that this reliance takes the form of an assumption that plays the role of an implicit premise in reasoning whose conclusion is the acceptance of the deliverance. Call this the *implicit premise thesis*. On ordinary assumptions, then, as soon as the implicit belief in the reliability of the faculty is put in question, one can hardly find support for that belief by appeal to the ostensible deliverances of that faculty. Such appeal would after all require a prior trust in the reliability of the faculty, and we would be in a vicious circle.

Is it perhaps on the basis of such reasoning that belief in one's own cognitive reliability is held importantly distinct from belief in the reliability as a

---

does not require such reflection (like the cognitio that Descartes allows to an atheist mathematician).

roller of a smooth stone, and even from belief in the cognitive reliability of someone else? In these other cases one can rationally base one's belief on one's perceptions, memories, and reasonings, overcoming thus any doubts that might be induced by the fact that the entity derives from brute forces. In one's own case it would be viciously circular to argue in parallel fashion from the deliverances of one's perception, memory, and reason to the reliability of these faculties seated in oneself.

How defensible is the *implicit premise thesis*? And, if not on the basis of that thesis, then how might we support the required distinction between R, on one hand, and the likes of S and F, on the other? How to support the claim that R is subject to defeat by brute origins, whereas S and F are exempt from such defeat? What accounts for this distinction if not the implicit premise thesis?

Something like the implicit premise thesis may be found in Thomas Reid, where it raises subtle and difficult issues. While unable to discuss these issues here, I do take them up elsewhere, where I reach the conclusion that no vicious circularity need be involved.[18] As a consequence, I can see no way to support the required distinction between R, on one hand, and S and F, on the other. So I do not believe that the foregoing line of argument can be successful.

IV

That brings us to the more promising reasoning in terms of the requirements of perspectival or reflective knowledge, reasoning which we now explore at least briefly: Does it offer better support for the case against naturalism?

Reflective knowledge requires an epistemic perspective underwriting the knower's belief from her own epistemic perspective. It might perhaps be argued that the naturalist, unlike the theist, is denied such a perspective, and is hence at an epistemic disadvantage. This, I submit, is actually the leading idea of Descartes's epistemological project. Descartes crafts for himself, and invites us to share, a theological perspective meant to underwrite his faculties and beliefs. If the faculties valued are those of a priori reflection, moreover, such as intuition and deduction, and if it is these one wants to cast in a good light, and if the best light is that attainable only by use of these very faculties—if these are "the most sure routes to knowledge, than which the mind should admit no others"—it is then hard to see what to invoke other than ra-

[18] I take up these issues in my discussion of Reid's epistemology, "Reidian First Principles," which is my part of a contribution on Reid, written jointly with James van Cleve, in *The Modern Philosophers: From Descartes to Nietzsche*, ed. Steven Emmanuel (Oxford: Blackwell, 1999). (Other aspects of Reid's philosophy are covered by James Van Cleve in the next chapter.)

tional theology. In fact it is easy to see why Descartes, with his background and in his milieu, would feel strongly drawn to that recourse.

Is the naturalist at an epistemic disadvantage? Is the naturalist indeed in a position that is rationally self-defeating? This seems implausible in light of the following. As a supernaturalist, one might use certain faculties (as does Descartes his faculties of a priori reflection) in devising an epistemically comforting view of one's universe, according to which one is so constituted and so related to one's world that one's faculties would be not misleading but truth-conducive. That I take to be Descartes's strategy.

What denies such a strategy to the naturalist? Why can't you as a naturalist develop a view of yourself and your surroundings that shows your situation to be epistemically propitious. You would need to be able to self-attribute ways of acquiring and retaining beliefs that put you so in touch with your surroundings, relative to the relevant fields of interest to you, that you would tend to believe correctly in virtue of using those ways. What precludes your doing so, by means of science, as a naturalist, if the supernaturalist can do so by means of theology?

And now comes a key question in assessing Plantinga's strategy: Is there any proposition O in a relevant sense just about our origins, whose inclusion in a view V would preclude V's enrichment by an epistemic perspective of the sort required for reflective knowledge. In other words, would any O-containing view V repel any possibility of being enriched coherently (while still including O) in a way enabling the holders of the enriched V to become thereby reflective knowers?

I myself can see no bright prospect for showing that the following would be such a proposition O: that we humans derive from brute forces. That proposition seems to me tenable compatibly with the view that in fact we are quite good cognizers of our surroundings through our use of vision, hearing, etc.; that these faculties put us reliably enough in touch with the truth about empirical and contingent goings-on around us, and that we are thereby enabled to know about these goings-on. Issues of circularity do arise as to how we can rationally and knowledgeably adopt such a view about our own epistemic prowess. But these problems of circularity are not exclusive to naturalism, as Descartes was soon to find out from his critics.[19]

[19] I myself think that Descartes was actually right on the issue of circularity, and the critics wrong, in ways specified in "How to Resolve the Pyrrhonian Problematic" and most recently in "Reflective Knowledge in the Best Circles," *Journal of Philosophy* 94 (1997): 410–430.

Can Atheists Know Anything?

James Van Cleve

Section 1 of this chapter explores solutions to the problem of the Cartesian Circle, briefly reviewing my own earlier work and focusing on recent articles by Ernest Sosa, Keith DeRose, and Mark Heller. Section 2 explores Plantinga's argument that naturalism (belief in evolution unguided by God) is epistemically self-defeating. In differing but related ways, Descartes and Plantinga both raise the question posed in my title.

I. Descartes's Answer

> I must inquire whether there is a God as soon as the occasion presents itself; and if I find that there is a God, I must also inquire whether He may be a deceiver; for without a knowledge of these two truths I do not see that I can ever be certain of anything.[1]

The problem of the Cartesian Circle arose for Descartes because he appeared to commit himself to each of the following two propositions:

(1) I can know (be certain) that (CD) whatever I perceive clearly and distinctly is true only if I first know (am certain) that (G) God exists and is not a deceiver.

[1] *The Philosophical Works of Descartes*, ed. Elizabeth S. Haldane and G. R. T. Ross (Cambridge: Cambridge University Press, 1911), 1:159. Hereafter I shall refer to the Haldane and Ross volumes as HR. I shall also be citing *The Philosophical Writings of Descartes*, trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch (Cambridge: Cambridge University Press, 1984), vols. 2 and 3, which I shall refer to as CSM.

(2) I can know (be certain) that (G) God exists and is not a deceiver only if I first know (am certain) that (CD) whatever I perceive clearly and distinctly is true.

Obviously, if (1) and (2) are both true, I can never be certain of either G or CD. To be certain of either, I would already have to be certain of the other. Yet Descartes said he was certain of both G and CD. How can this be possible?

Any adequate solution to the problem of the Cartesian Circle must plainly deny either (1) or (2). I begin by reviewing a solution I proposed in 1979, which denies (2).[2]

A. A Purely Externalist Solution

The key to my solution is a distinction between the following two propositions, either of which might be expressed by the ambiguous sentence 'I am certain of the truth of clear and distinct perceptions':

(A) For any p, if I clearly and distinctly perceive that p, then I am certain that p.
(B) I am certain that: for any p, if I clearly and distinctly perceive that p, then p.

The chief difference between (A) and (B) is that (A) says whenever I clearly and distinctly perceive any proposition I will be certain of *it* (the proposition in question), whereas (B) says that I am certain of a *general principle* connecting clear and distinct perception with truth. Clearly, (A) could be true of someone even though (B) were not.

Such, I suggest, is precisely Descartes's situation before he has proved the existence and veracity of God. He is not then certain that all clear and distinct perceptions are true, but that does not prevent him from becoming certain of the various particular propositions on which the light of clear and distinct perception falls. These propositions include the *cogito* as well as the various premises he uses in proving the existence and veracity of God. Once he has proved the latter, he obtains as a corollary the principle CD—that whatever he perceives clearly and distinctly is true.[3] But nowhere in this chain of reasoning does he need to use CD as a premise or know that it is true. Therefore, proposition (2) above is false, and the circle is broken.

The strategy I impute to Descartes is in two important ways externalist, for it gives an answer of *no* to each of the following questions:

(i) In order for knowledge-conferring factors to do their job, must the subject know that they obtain?

[2] James Van Cleve, "Foundationalism, Epistemic Principles, and the Cartesian Circle," *Philosophical Review* 88 (1979): 55–91.

[3] Perhaps he can prove that clear and distinct perceptions are not only true but certain, in which case he would be proving proposition (A) itself.

(ii) Must the subject also know something about the "justificatory efficacy" of such factors?[4] For example, must he know that the factors in question do confer knowledge, or that they are a reliable source of true beliefs?

The prime knowledge-conferring factor for Descartes is, of course, clear and distinct perception. As I see things, clear and distinct perception can give knowledge even to a subject who does not know that he is in that state. Thus Descartes says no to question (i).[5] More importantly, he also says no to question (ii). Clear and distinct perception can give knowledge even to a subject who knows nothing about the epistemic powers of such perception—nothing about its reliability or its justificatory efficacy. It is externalism on this second score that is crucial for escaping the circle.

I turn now to alternative interpretations of Descartes that have been put forth by Sosa, DeRose, and Heller.[6] The differences among the four of us can be summed up as follows. According to Sosa and DeRose, simple clear and distinct perception has *less* power than I say it does, for what it delivers is not knowledge of the highest grade. According to Heller, on the other hand, clear and distinct perception has *more* power than I say it does, for it delivers not only knowledge of the highest grade but also knowledge that one has this knowledge and knowledge of the reliability of clear and distinct perception itself.

B. The Two-Levels Solutions of Sosa and DeRose

Common to the solutions of Sosa and DeRose is a distinction between two kinds or levels of knowledge, marked in Descartes by the terms *cognitio* and *scientia*. To prepare the way for their positive views, I shall first discuss two passages they cite that make trouble for my interpretation.

The first is the notorious sentence occurring at the end of the fourth paragraph in the Third Meditation and serving above as my epigraph:

> I must inquire whether there is a God as soon as the occasion presents itself; and if I find that there is a God, I must also inquire whether He may be a deceiver;

[4] I take the phrase in quotation marks from William P. Alston, "An Internalist Externalism," *Synthese* 74 (1988): 265–283.

[5] Descartes may well hold that we *can* know, just by reflection or introspection, that our state is one of clear and distinct perception, in which case he would be a "weak internalist" in BonJour's sense and an "access internalist" in Alston's. What I am saying here is that he is not a "strong internalist" in BonJour's sense. See William Alston, "Internalism and Externalism in Epistemology," *Philosophical Topics* 14 (1986): 179–221, and Laurence BonJour, "Externalism/Internalism," in *A Companion to Epistemology*, ed. J. Dancy and E. Sosa (Oxford: Blackwell, 1992), 132–136.

[6] Keith DeRose, "Descartes, Epistemic Principles, Epistemic Circularity, and Scientia," *Pacific Philosophical Quarterly* 73 (1992): 220–238; Ernest Sosa, "How to Resolve the Pyrrhonian Problematic: A Lesson from Descartes," *Philosophical Studies* 85 (1997): 229–249; Mark Heller, "Painted Mules and the Cartesian Circle," *Canadian Journal of Philosophy* 26 (1996): 29–56.

> for without a knowledge of these two truths I do not see that I can ever be certain of anything.[7]

In saying that proposition (A) is true of Descartes all along, I had been saying that clear and distinct perception suffices to make him certain of things. Here he is apparently saying that clear and distinct perception does *not* suffice—that in addition he must be sure that God exists and is no deceiver. So if we take Descartes at his word here, proposition (A) is not true of him at this stage of the Meditations.

It is tempting to sweep this sentence under the rug, since it apparently creates insuperable difficulties in Descartes's ever attaining any knowledge at all. The prima facie difficulties are well brought out by Sosa, who calls our attention to the form of the problematic sentence "I cannot know anything unless I know that p." If all the rest of my knowledge depends on my knowledge of one special fact, I had better be able to know that special fact by immediate intuition—if I had to know it through the mediation of other items of knowledge, I could never know it all. Yet Descartes does not hold that God's existence is known by immediate intuition. So Sosa observes, "It seems incoherent for Descartes to contend that he cannot know anything unless he knows there is a veracious God."[8]

The second passage that apparently makes trouble for my interpretation is one in which Descartes comments on the epistemic advantage he has over the atheist. One of the authors of the Second Objections had asked: may not an atheist know that the three angles of a triangle are equal to two right angles, despite not knowing (and indeed denying) that there is a God?[9] To this Descartes made the following reply:

> That *an atheist can know clearly [clare cognoscere] that the three angles of a triangle are equal to two right angles*, I do not deny; I merely affirm that, on the other hand, such knowledge on his part [*cognitionem*] cannot constitute true science [*scientia*], because no knowledge that can be rendered doubtful should be called science. Since he is, as supposed, an Atheist, he cannot be sure that he is not deceived in the things that seem most evident to him, as has been sufficiently shown; and though perchance the doubt does not occur to him, nevertheless it may come up, if he examines the matter, or if another suggests it; he can never be safe from it unless he first recognizes the existence of a God.[10]

[7] HR 1:159; for an alternative translation, see CSM 2:25.

[8] Sosa, "How to Resolve the Pyrrhonian Problematic," 235. At p. 238, however, Sosa says that Descartes's claim is incoherent only on the assumption that Descartes is an unqualified foundationalist, rather than a two-levels theorist as Sosa takes him to be.

[9] HR 2:26; CSM 2:89. The same objector notes that if Descartes was serious in saying that he knows nothing until he knows there is a God, then he did not even know that he was a thinking thing in the Second Meditation.

[10] HR 2:39; I shall be remarking below on my reasons for preferring this translation to that of CSM 2:101.

According to my account,[11] the atheist can be as certain as Descartes that the angles of a triangle are equal to two right angles if he is clearly and distinctly perceiving this fact at the time. His disadvantage comes out only at later times when neither of the two is attending to the proof of the theorem, but only remembering having clearly and distinctly perceived its truth before. At such times Descartes will retain knowledge of the theorem, but the atheist will not. This is because Descartes, having obtained a divine guarantee of the truth of clear and distinct perceptions, is in a position to draw the inference from 'I formerly had a clear and distinct perception of p' to 'p is true', whereas the atheist is not. But (to repeat) at times when Descartes and the atheist both have a clear and distinct perception of a given proposition, they will be epistemic equals.[12]

With his eye on the above-quoted passage, however, Sosa makes the following observation to the contrary:

> Here Descartes is not claiming only *ex post facto* advantage. Take the moment when both are clearly and distinctly perceiving the fact that the three angles are equal to two right ones. *Even at that very moment*, according to Descartes, the atheist is at an epistemic disadvantage.[13]

What is the epistemic advantage that Descartes possesses over the atheist even at that very moment?

One can agree that there is at least this much difference between Descartes and the atheist at time t: the atheist's knowledge is loseable and Descartes's (once he is fortified with knowledge of God) is not.[14] But the "difference at that very moment" Sosa has in mind is more robust than this.

It is explicit in our passage that Descartes and the atheist both possess cognitio at the moment of clear and distinct perception, but that Descartes alone possesses scientia. What epistemological difference lies behind these words? Although Descartes never elaborates the distinction in any systematic fashion, DeRose and Sosa each develop the distinction in their own ways and make it the key to their interpretations of Descartes.

DeRose offers no analysis of the *meaning* of 'scientia', but suggests the fol-

[11] Van Cleve, "Foundationalism, Epistemic Principles, and the Cartesian Circle," 68.

[12] Substantial textual support for this interpretation may be found in the following places: (i) the next-to-last two paragraphs of the Fifth Meditation (HR 1:184–185; CSM 2:48–49), (ii) Descartes's response to the third objection in the second set of objections to the Meditations (HR 2:38, 39; CSM 2:100–101), (iii) his response to Arnauld's charge of circularity in the fourth set of objections (HR 2:114–115; CSM 2:171), (iv) Descartes's Conversation with Burman, trans. John Cottingham (Oxford: Clarendon Press, 1976), 6, and (v) Descartes's letter to Regius, CSM 3:147.

[13] Sosa, "How to Resolve the Pyrrhonian Problematic," 236.

[14] Moreover, this is a difference at t in more than merely a Cambridge sense, such as would be constituted by the fact that it is true at t that the atheist will lose his knowledge at t + 1 whereas Descartes will not. It is a difference in the dispositional features of their respective states of knowledge, features possessed at t even if they are manifested only in future or hypothetical circumstances. The letter to Regius cited in note 12 makes this point.

lowing as giving necessary and sufficient conditions for it: S has scientia of p if and only if (1) S clearly and distinctly perceives that p is true, and (2) S clearly and distinctly perceives the truth of the principle that what S clearly and distinctly perceives is true.[15] Given that clear and distinct perception is the only source of cognitio for Descartes, these conditions are equivalent to the following: S has scientia of p iff S has cognitio of p plus cognitio of the reliability of the original cognitio's source.

Why is the atheist, who possesses cognitio but not scientia of the theorem, in an inferior position? According to DeRose, the answer is that the atheist is *vulnerable to skeptical attack*.[16] The thought that there may be an arch-deceiver who is able and willing to make even clear and distinct perceptions false serves to undermine his belief in the theorem. By contrast, once Descartes has proved the existence of a veracious God and drawn the corollary that all clear and distinct perceptions are true, he is in a position to turn aside such skeptical attack.

What DeRose says may be put in the following terms: the atheist (but not Descartes) is vulnerable to confrontation with a *defeater* that he cannot dispel.[17] So stated, what DeRose says invites comparison with Plantinga's argument against the naturalist, which I shall discuss in section 2 of this chapter.

Is DeRose saying that the cognitio normally conferred by clear and distinct perception may be destroyed even at the time of clear and distinct perception by the consideration of defeating hypotheses? At one place he seems to say no: *cognitiones* are not discarded in the face defeating doubts but are merely prevented from rising to the higher epistemic level of scientia.[18] On the other hand, what he says about skeptical hypotheses as *undermining* their target beliefs suggests that cognitio is, or ought to be, destroyed.[19] The normal idea of a defeater is that (unless defeated itself) it takes away whatever element of initial warrant there was in the belief it defeats.

Be that as it may, I have the following misgiving about DeRose's account. If the hypothesis of a deceiving God serves as a defeater for anything the atheist believes on the strength of clear and distinct perception, does it not serve likewise as a defeater for Descartes's belief in God? For that belief, too, is based on clear and distinct perception. DeRose's account of the atheist's disadvantage threatens to make it a crippling disability for everyone and a victory only for the skeptic.

I turn now to Sosa's account. Sosa suggests that Descartes's distinction between cognitio and scientia may be seen as paralleling the distinction he has been developing for some years now between animal knowledge and reflec-

[15] DeRose, "Descartes, Epistemic Principles," 224.

[16] Ibid., 228.

[17] DeRose speaks instead of hypotheses that "undermine" their target beliefs, but the idea of a defeater is plainly there.

[18] DeRose, "Descartes, Epistemic Principles," 233–234.

[19] Ibid., 226–227.

tive knowledge. Animal knowledge is (roughly) reliable true belief,[20] such as unreflective beasts may have and such as even a thermometer might have if it could be said to have beliefs at all. Reflective knowledge, on the other hand, is a meta-state that arises only once one has achieved a perspective on one's own beliefs and their reliability.

Reflective knowledge is governed by the following *principle of epistemic ascent*: if one knows that p and understands what it is to know that p, then one implicitly believes that one knows that p and is justified in this belief. In symbols, Kp & UKp ⇒ JKp. Sosa combines this principle with another principle, the principle of transfer of justification through justifiably believed entailment, JX & J(X ⇒ Y) ⇒ JY, to obtain two significant corollaries that also govern reflective knowledge.[21] The first corollary is the *principle of exclusion*, Kp & J(Kp ⇒ ~q) ⇒ J~q, which says that if you know something and are justified in believing that your knowing it excludes q, you are also justified in believing q to be false. This principle may also be derived in positive form: if one knows something and is justified in believing that one's knowing that thing requires q, then one must also be justified in believing q to be true. Now one of the requirements of one's knowing that p is that the sources of one's belief in p be minimally reliable; furthermore, it is arguable that this entailment is justifiably believed in by anyone who understands what knowledge is: UKp ⇒ J(Kp ⇒ Rp). Putting all this together, we arrive at the second corollary, which Sosa calls the *principle of the criterion*: Kp & UKp ⇒ JRp. This says that if you know something and understand what knowledge is, then you are justified in believing that the source of your belief is reliable.

I wish to point out that by employing somewhat strengthened forms of the premises Sosa uses, we may derive in analogous fashion the following *strong principle of the criterion*: Kp & UKp ⇒ KRp.[22] That is, anyone who knows any proposition p and knows what knowledge is must know that the source of his belief in p is reliable.

The principle of the criterion, especially in its strong form, may look at first as though it sets up an even bigger obstacle to knowledge than Descartes's notorious sentence in the Third Meditation. Like Descartes's sentence, it implies that one's knowledge of any proposition q depends on one's knowledge of a special proposition p—in this case a proposition to the effect that the sources of one's belief in q are reliable.[23] But this time not even im-

[20] For present purposes, I abstract from the differences between Sosa's "virtue" account of animal knowledge and a straight reliabilist account.

[21] For details of the derivations, see Sosa, "How to Resolve the Pyrrhonian Problematic," 232–233, 243–245.

[22] The proof of the strong principle requires epistemic ascent in the strengthened form Kp & UKp ⇒ KKp (which Sosa says he would defend [ibid., 246 n. 15]). It also requires the principle that knowledge is transferred by known entailment and the premise that understanding what knowledge is involves knowing that it requires reliability.

[23] Insofar as different beliefs may have different sources, it need not be the same proposition p in all cases.

mediate knowledge of the special proposition could save the day, since the strong principle of the criterion arguably implies that there is no immediate knowledge—it makes every bit of knowledge depend on a further bit.

How are we to surmount this obstacle—how are we to obtain the knowledge of reliability that is a condition of knowing (in the reflective way) anything whatever? The parenthetical qualification points the way toward the answer. The knowledge of reliability that boosts us up to reflective knowledge may itself be compounded out of animal knowledge. As Sosa notes elsewhere, "Human reflective knowledge is most likely to depend ultimately on unreflective knowledge, since we cannot climb infinite ladders of reflection."[24]

The suggestion, then, is that by amassing enough animal knowledge, we may eventually attain reflective knowledge. In the Cartesian setting, things might proceed like this: First, we attain various object-level cognitiones by letting the light of clear and distinct perception fall on various truths. From these first truths we go on to infer various other things, including most importantly the existence and veracity of God. From that in turn we infer that all clear and distinct perceptions are true. As soon as we have taken that last step, all of our originally animal cognitiones become pieces of reflective scientia, for they are now augmented by cognitiones of the reliability of their source.[25]

In the procedure just sketched, we give an argument for the reliability of clear and distinct perception using premises and inferences authorized by nothing other than clear and distinct perception itself—we argue for the reliability of a faculty using deliverances of that very faculty. Is there anything objectionably circular about such arguments? I think not. On the assumption that clear and distinct perception gives knowledge only if it is reliable, such arguments do indisputably have the feature that William Alston has chosen to call 'epistemic circularity': the truth of the argument's conclusion is a necessary precondition of obtaining knowledge of the argument's premises.[26] As DeRose points out, however, Alston is using 'epistemically circular' merely as a label for this type of argument.[27] In my opinion, it is not a label to which we should react as we do to the skull-and-crossbones label on a bottle of poison. The really damaging kind of circularity, the kind to be shunned, is the kind in which not merely the *truth* of an argument's conclusion, but *knowledge* of it, is a necessary precondition of obtaining knowledge of the argument's premises. The Cartesian argument sketched above is not circular in that way, for to re-

[24] Ernest Sosa, *Knowledge in Perspective* (Cambridge: Cambridge University Press, 1991), 290.

[25] We would presumably also need, for each of the original cognitiones, a cognitio that its source was clear and distinct perception.

[26] See William Alston, "Epistemic Circularity," *Philosophy and Phenomenological Research* 47 (1986): 1–30.

[27] DeRose, "Descartes, Epistemic Principles," 230.

peat: clear and distinct perception gives us knowledge (if only at the cognitio level) even if we do not yet know that clear and distinct perception is reliable.

Critical assessment of the two-levels solution (which I find quite attractive) must turn in good measure on why scientia is superior to cognitio—a point on which DeRose and Sosa both have instructive things to say, but into which I shall not enter here. Instead, I wish simply to note two points that I think the two-levels solution must retain from the purely externalist solution.

First, looking back to propositions (1) and (2) at the beginning, how are we to break the circle? Sosa notes that if the epistemic terms in (1) and (2) are appropriately disambiguated, Descartes may affirm versions of *both* (1) and (2): (1') there is scientia of CD only if there is first cognitio of G, and (2') there is scientia of G only if there is first cognitio of CD.[28] I agree. But I note as well that if we read the epistemic terms in (1) and (2) as 'cognitio' throughout, we must deny one of them or face epistemic disaster. Sosa, DeRose, and I would all agree that the proposition to be denied in that case is (2). That is, we can have cognitio-knowledge through clear and distinct perception without knowing that such perception is reliable. At least at the level of cognitio, externalism must be upheld.

Second, it is crucial to the Cartesian enterprise (as DeRose notes)[29] that the lower of the two epistemic levels (cognitio) already be very high. I would say, in fact, that it must be high enough not to be subject to defeaters in the way DeRose envisions. If Descartes's cognitio of God can knock out defeaters, why is his original cognitio subject to them? I think cognitio must be capable of standing on its own as genuine knowledge if the scientia compounded out of it is to be knowledge of a superior sort.[30]

C. The Internalist Solution of Heller

Mark Heller has put forth an interpretation of Descartes that has interesting points of agreement and disagreement with the interpretations canvassed above.[31] Heller holds that the knowledge possessed by an atheist at a moment when he is clearly and distinctly perceiving the truth of some geometrical proposition is as good as knowledge ever gets. Proving the existence of God does nothing to make Descartes's knowledge superior to that of the atheist; Descartes's advantage emerges only at times when the atheist is no longer en-

[28] Similar disambiguation helps take the sting out of Descartes's notorious sentence: "without a knowledge [cognitio] of these two truths I do not see that I can ever be certain [have scientia] of anything" (HR 1:59).

[29] DeRose, "Descartes, Epistemic Principles," 228, 231–232.

[30] For this reason I prefer the HR translation of *cognitio* and *scientia* as 'knowledge' and 'science' respectively to the CSM translation of them as 'awareness' and 'knowledge'. Perhaps lower-level knowledge may be parlayed into higher-level knowledge, but how can nonknowledge be parlayed into knowledge?

[31] Heller, "Painted Mules and the Cartesian Circle," 29–56.

joying clear and distinct perception. On this point Heller implicitly sides with my interpretation and against that of Sosa and DeRose. Yet Heller brings out a way in which the atheist's knowledge is precarious that I had not brought out. As I presented things, the atheist who merely remembers that he once clearly and distinctly perceived p, but lacks Descartes's knowledge that all clear and distinct perceptions are true, has no basis for concluding that p. This may make it appear that the atheist's problem is simply a lost basis for sustaining belief once clear and distinct perception has faded away. Heller points out that even if the atheist retains his belief in former objects of clear and distinct perception, he will be vulnerable to skeptical challenge in a way that Descartes is not.

The skeptical challenge Heller has in mind takes the form not of a defeater (as in DeRose's account), but of an argument based on a closure principle: if you know p and you also know that p entails q, then you know q. A typical such argument runs as follows:

1. If I know that 2 + 3 = 5, then I know that no evil demon is making me believe falsely that 2 + 3 = 5. (Given our closure principle, this premise will be true of anyone who knows that '2 + 3 = 5' entails 'no evil demon is making me believe falsely that 2 + 3 = 5'.)
2. I do not know that there is no such demon.
3. Therefore, I do not know that 2 + 3 = 5.

Recent philosophy has recognized three main responses to such arguments. One may acquiesce in the conclusion—that is the skeptical response. One may deny the first premise by denying its underlying closure principle—that is the "Nozickian" response. Finally, one may make bold to deny the second premise—that is the "Moorean" response.[32]

According to Heller, if the atheist is clearly and distinctly perceiving that 2 + 3 = 5 at the time when the above challenge is issued, he is fully entitled to make the Moorean response. If he is not then clearly and distinctly perceiving the proposition, he has no response.[33] Under the same circumstances, however, Descartes does have a response, for having proved the existence of a veracious God who would not allow the demon to operate, he may still reject the second premise.[34]

[32] Recent epistemology has also produced a fourth option, the "contextualist" response. For elucidation and defense, see Keith DeRose, "Solving the Skeptical Problem," *Philosophical Review* 104 (1995): 1–52.

[33] Perhaps 2 + 3 = 5 is one of those immediately obvious propositions that one cannot entertain without having a clear and distinct perception of it, in which case anyone who entertained the skeptical argument would automatically be in a position to give the Moorean response to it. If so, we should change the proposition in Heller's example to one that is not so immediately obvious.

[34] This presupposes that Descartes may rely on his memory. If the same privilege is extended to the atheist, I fail to see that Descartes has much of an advantage. On this point, see Van Cleve, "Foundationalism, Epistemic Principles, and the Cartesian Circle," 68 n. 31.

How is it that present clear and distinct perception entitles one to make the Moorean response? I have the impression that Heller would answer along the following lines: the force of clear and distinct perception is *absolutely overwhelming*. It is so irresistible that it produces a kind of knowledge that penetrates to all known consequences of what is known (i.e., to all propositions known by clear and distinct perception to follow from what is likewise known.) The point is not just that knowledge by clear and distinct perception is closed under clearly and distinctly known entailment—that much the skeptic accepts—but also that Moore's *modus ponens* trumps the skeptic's *modus tollens*.

I said above that whereas DeRose and Sosa attribute less power to clear and distinct perception than I had attributed to it, Heller attributes more. The reader may surmise that the "more" consists in the overwhelming force I have just mentioned, but that is not what I have in mind.[35] The "more" makes its appearance in the following two claims:

> This further possibility [of some arithmetical proposition's being false despite being clearly and distinctly perceived] is eliminated by the higher order knowledge produced by clear and distinct perception—Descartes knows that, whatever mechanism it is that is producing the arithmetic belief, it is infallible.[36]

and

> Clear and distinct perception of p produces absolute certainty not only in p, but in the proposition that the subject knows p, and that she knows that she knows p, and so on for every epistemic level.[37]

Identified here are two types of metaknowledge: knowledge about the epistemic powers of clear and distinct perception and knowledge that one knows. For my Descartes, such metaknowledge is achieved only after one proves the existence of God, and it is optional so far as object-level knowledge is concerned. For Heller's Descartes, by contrast, both types of metaknowledge are produced by clear and distinct perception whenever it produces any knowledge at all.

It is instructive to compare Heller on this score with Sosa. Like Sosa, Heller attaches considerable importance to the features distinctive of reflective knowledge—its being accompanied (necessarily) by knowledge of the reliability of its source and (typically) by knowledge that one knows. But unlike Sosa, he does not see these features as being animal and gradual in their origin; they are automatic deliverances of clear and distinct perception itself. Heller even makes the intriguing suggestion that Descartes is an *internalist reliabilist*: reliabilist because knowledge must arise from a reliable (indeed infal-

[35] I am not averse to attributing such force to clear and distinct perception; see Van Cleve, "Foundationalism, Epistemic Principles, and the Cartesian Circle," 65–66.

[36] Heller, "Painted Mules and the Cartesian Circle," 45.

[37] Ibid.

lible) mechanism, internalist because clear and distinct perception might be "an introspective mechanism that would indicate its own reliability every time it was used."[38]

Reading what Heller proposes, I can almost believe that I am hearing the ghost of that old arch-internalist, H. A. Prichard. Three distinct Prichardian themes are echoed in Heller. First, there is the use of closure principles to dispose of skeptical hypotheses:

> In knowing that a triangle must have such angles we also know that nothing can exist which is incompatible with this fact.[39]

The "Moorean" response might with equal justice have been called the "Prichardian" response.

Second, there is an internalist version of the KK-thesis:

> We must recognize that whenever we know something we either do, or at least can, by reflecting, directly know that we are knowing it, and that whenever we believe something, we similarly either do or can directly know that we are believing it and not knowing it.[40]

> We therefore can get Descartes out of the impasse, provided we allow, *as we can*, that the state to which he referred as one of perceiving clearly was in fact one of knowing, and one which he could have known at the time to be one of knowing. By doing this we stop the rot from starting. Descartes's trouble was that he let the rot begin, and once it is allowed to start, it cannot be prevented from going on indefinitely.[41]

In the second of these passages, we find Prichard not only invoking a version of the KK-thesis but suggesting that it is indispensable for warding off skeptical attack.

Third, there is the internal accessibility of reliability. Prichard never wrote the following words, but I can easily imagine him proclaiming them:

> When I reflect on my current condition—of clearly and distinctly perceiving p—I see that I could not be related to any proposition as I am related now to p unless that proposition were true.

[38] Ibid., 50. I am not sure how introspective knowledge of the reliability of clear and distinct perception is meant to relate to Heller's account of knowledge of such reliability on pp. 30–31.

[39] H. A. Prichard, "Descartes's Meditations," in *Descartes*, ed. Willis Doney (Garden City, N.Y.: Doubleday Anchor Books, 1967), 151. What Prichard says here could be symbolized as $Kp \rightarrow K{\sim}\exists q(q \,\&\, q \Rightarrow {\sim}p)$ or (substituting an equivalent) $Kp \rightarrow K[(q)(q \Rightarrow {\sim}p) \rightarrow {\sim}q)]$. If we move the second K inside the quantifier and distribute it across the conditional, we obtain $Kp \rightarrow [(q)K(q \Rightarrow {\sim}p) \rightarrow K{\sim}q)]$; if contraposition is acceptable within K-contexts, the latter formula is equivalent to closure under known entailment.

[40] Prichard, "Descartes's Meditations," 147.

[41] Ibid., 152.

Indeed, he would be committed to such a position if he accepted a reliability constraint on knowledge alongside his version of the KK-thesis.

I would like to submit for the reader's consideration the possibility that we can we pry apart the first of the Prichardian elements in Heller's account from the other two. Might clear and distinct perception have enough epistemic force to warrant the Moorean response to skeptical arguments while not being automatically reflective in the two ways Heller supposes? If so, we could agree with Heller that anyone in a state of clearly and distinctly perceiving p has knowledge of p that withstands skeptical challenge, even if the subject does not yet know that he knows p or that the cognitive mechanism generating his belief in p is reliable.

D. Summary

Can atheists know anything? The answers given by Descartes according to the three interpretations we have considered here are as follows.

The externalist interpretation (Van Cleve): yes. Knowledge of God may be required for knowledge of the reliability of our faculties, for higher-order knowledge (knowledge that one knows), and for knowledge of certain items that are no longer clearly and distinctly perceived. But it is not required for knowledge at the object level; for that, clear and distinct perception of the object-level proposition suffices by itself. Moreover, the knowledge afforded by simple clear and distinct perception is as good as knowledge gets.

The two-levels interpretation (Sosa and DeRose): yes and no. Persons ignorant of God may enjoy the lower grade of knowledge, cognitio, for which clear and distinct perception suffices as above. However, there is a higher grade of knowledge, scientia, to which one ascends only upon coming to know that one's faculties are reliable, and such knowledge must be grounded for Descartes in knowledge of a veracious God.[42]

The internalist interpretation (Heller): yes. All knowledge whatever must satisfy the higher-level demands of scientia—in particular, it must be backed by knowledge of the reliability of its source. But for knowledge of the reliability of clear and distinct perception at the time that one is having it, no knowledge of God is required—all it takes is introspection upon the workings of one's own mind.

II. Plantinga's Answer

> In whatever way they [those who deny the existence of an all-powerful God] suppose that I have arrived at the state of being that I have

[42] I have not addressed here the question whether knowledge of the reliability of one's faculties is obtainable only through knowledge of God.

> reached—whether they attribute it to fate or to accident, or make out that it is by a continual succession of antecedents, or by some other method—since to err and deceive oneself is a defect, it is clear that the greater will be the probability of my being so imperfect as to deceive myself ever, as is the Author to whom they assign my origin the less powerful.[43]

In what could almost be an extended commentary on Descartes's offhand remark, Alvin Plantinga has assembled an intricate and challenging argument to the effect that belief in evolution unguided by God undermines its own rationality.[44] To encapsulate the argument, let R be the proposition that human cognitive faculties are reliable; let E be the proposition that human beings are the products of evolution; and let N be naturalism, a view denying the existence of God or any supernatural beings, so that evolution, if it occurred, was not orchestrated by God. Plantinga argues that the probability of R given the conjunction N&E is either low or inscrutable. He argues further that this fact gives the believer in N&E a defeater for R, making continued belief in R on his part irrational. He argues finally that once the naturalist withholds belief in R, he has a defeater for all of his beliefs, including N&E itself. The conclusion is that naturalism—or rather the combination of naturalism with evolution, but what naturalist does not believe in evolution?—is self-defeating.

My interest in this argument is not in its final step, the allegedly self-defeating character of naturalism, but in the penultimate step, in which Plantinga asserts that the naturalist has a defeater for everything he believes. If so, then on the assumption that no defeated belief can amount to knowledge, it follows that the naturalist has no knowledge whatsoever. It appears that Plantinga renders a more severe verdict on the atheist's epistemic position than Descartes ever did: Descartes allowed that the atheist geometer might know (for a while, or in a way) a simple theorem or two, but Plantinga's naturalist must evidently be a complete ignoramus.

In what follows I will concede for the sake of argument that the probability of R on N&E is at best inscrutable—a point for which Plantinga offers formidable defense. The issues I will raise are (1) whether this gives the naturalist a defeater (or more accurately, an undefeatable or ultimately undefeated defeater) for R, (2) whether that in turn would give him a defeater for everything else he believes, thereby depriving him of all knowledge, and (3) whether, if the naturalist's predicament is as bad as Plantinga says it is, the theist's situation is any better.

Before proceeding we must say something about the concept of a defeater, which plays such a crucial role in Plantinga's argument. A rough and ready gloss of Plantinga's account is the following: proposition D would defeat be-

[43] HR 1:147.

[44] Alvin Plantinga, *Warrant and Proper Function* (New York: Oxford University Press, 1993), chap. 12, hereafter *WPF*; and "Naturalism Defeated" (unpublished).

lief B for a subject with noetic structure N (specified by describing all the subject's beliefs and experiences) iff one with N who came to believe D could not rationally continue to believe B.[45] Defeaters may themselves be defeated, which is to say that even though D defeats B for someone with N, D may no longer defeat B for someone whose noetic structure is N augmented by appropriate further beliefs. Note, then, that a "defeater-defeater" need not make belief in the original defeater irrational; it may instead make it rational to retain the initially defeated belief alongside the original defeater of it.[46] A defeater is ultimately undefeated if it is either undefeated or defeated by no beliefs that are not themselves defeated no matter how the subject's noetic structure is enriched.

A. Does the Naturalist Have a Defeater for R?

In an article refining his argument and defending it against various critics, Plantinga surveys and addresses five objections that bear on my first question: does the naturalist have a defeater for R? My own critique will incorporate three of these objections; as will emerge, I think they work better jointly than separately.

Objection 1: Plantinga states the first objection, which he dubs "the perspiration objection," as follows:

> You claim that the naturalist has a defeater for R in the fact that the probability of R on N&E is either low or inscrutable. But this can't be right. The probability that the function of perspiration is to cool the body, given (just) N&E, is also low, as is the probability that Holland, Michigan, is thirty miles from Grand Rapids, given N&E. But surely it would be absurd to claim that these facts give the partisan of N&E a defeater for those beliefs.

As Plantinga notes, what the objector is finding at work in his argument is the following general principle, which I shall label *Principle 1*:

> For any propositions A and B that I believe, if B is improbable or inscrutable with respect to A . . . , then A is a defeater for B.[47]

This is the principle Plantinga appears to be invoking in claiming that R's low or inscrutable probability on N&E gives the naturalist a defeater for R. The objector's point is that the principle is surely false.

I agree, but I would give the objection a different twist. In the version

[45] See "Naturalism Defeated," 32–33, for a more elaborate definition plus commentary.

[46] In the terminology of "Naturalism Defeated," 35, there are neutralizing defeater-defeaters as well as intrinsic defeater-defeaters.

[47] "Naturalism Defeated," 39. For brevity's sake, I will follow Plantinga in speaking of A as a defeater for B, even though the real defeater (as Plantinga notes on pp. 13 and 36–37) is not A alone, but A together with the subject's recognition of its epistemic relation to B.

above, the objector plugs the "perspiration proposition" into slot B in Principle 1, letting it play the role of defeatee. I believe we get a stronger objection if we plug the perspiration proposition (or whatever "innocent" proposition we choose for our example) into slot A, letting it be the defeater.

Once we make the switch, the objection takes the following form. For anything B that I believe, there is almost sure to be something else A that I believe such that B has inscrutable probability given A.[48] I believe that it is raining now; I also believe that the Broncos won Super Bowl XXXII; and I have no idea what the probability of the first is given the second. Does my Super Bowl belief therefore defeat my belief about the rain? If Principle 1 is correct, the answer is yes.[49] That might not be so bad if it were an easily defeated defeater. Yet by Plantinga's reasoning in the case where R is alleged to be initially defeated by N&E, it will be an undefeatable defeater. To neutralize the defeating force of the Super Bowl proposition, I might try appealing to other items of evidence for the rain—the splashes I see in the puddles outside, or the patter I hear on the roof. But such appeal will be to no avail if Principle 1 is correct, for my beliefs in such corroborating phenomena themselves have inscrutable probability in relation to the Super Bowl proposition, and are therefore defeated by it. Plantinga does not allow any proposition to serve as a defeater-defeater if it is already defeated itself.

Plantinga's response to the perspiration objection is to repudiate the principle against which it is directed. Indeed, he cites a counterexample of his own to Principle 1—a case in which the alleged defeater derives its warrant (or is rationally believed by the subject to derive its warrant) from the alleged defeatee. However, he does not tell us (so far as I can see) what principle he would substitute for Principle 1 in his argument. And if we amend Principle 1 to take care of the only difficulty he notes in it, we still get a principle that causes trouble for everyone, not just the naturalist. Amending the principle specifically to avoid Plantinga's counterexample, we would obtain this:

> For any propositions A and B that I believe, if B is improbable or inscrutable with respect to A, then A is a defeater for B unless A derives its warrant from B.

[48] This is illustrated by the second objection Plantinga considers, the "austere theism" objection, according to which the theist, too, believes something (namely austere theism) on which R has low or inscrutable probability.

[49] In the chapter in *WPF* on epistemic probability, Plantinga makes the following observation: "I might know a pair of propositions, one of which was epistemically improbable with respect to the other. I know, for example, both 99 out of 100 Americans are less than 6 foot 4 inches tall and Harry is an American, and Harry is 6 foot 5 inches tall" (143). Does this not fly in the face of Principle 1? Not necessarily. Plantinga could say that the first proposition in the pair about Harry does indeed defeat the second, but that one can nonetheless know them both—e.g., by having independent evidence for the second that neutralizes the defeating force of the first. That, of course, raises the question whether the defeating force of N&E can similarly be neutralized, which will occupy us a good deal in what follows.

The Super Bowl proposition does not derive its warrant from what I believe about the rain, so according to the revised principle, it still serves as a defeater for it—thus apparently depriving me of any knowledge of the rain.

To see just how dangerous Principle 1 really is, note that R itself presumably has inscrutable probability given the Super Bowl proposition,[50] and would therefore (according to Principle 1) be defeated by it. But once R is defeated, reasoning parallel to Plantinga's against the naturalist would show that *everything* one believes is defeated, making the situation irredeemable. Any premises used to bolster R or neutralize the defeating force of the Super Bowl proposition would include propositions some of which (I wager) have inscrutable probability with respect to the Super Bowl proposition; by Plantinga's reasoning, they would therefore be defeated by it. Witness:

> For consider any argument from science (or anywhere else) [the naturalist] might produce [by way of saving R]. This argument will have premises; and these premises, he claims, give him good reason to believe R. . . . But note that he has the very same defeater for each of those premises that he has for R. . . . [F]or each of the premises, he has a reason for doubting it, a reason for being agnostic with respect to it. This reason, obviously, cannot be defeated by an ultimately undefeated defeater. For every defeater of this reason he might have, he knows that he has a defeater-defeater: the very undercutting defeater that attached itself to R in the first place.[51]

Echoing Prichard, we might put Plantinga's point by saying that defeat of R is a rot that can't be stopped once it has been allowed to start. My complaint against Principle 1 is that it lets the rot begin.[52]

Objection 3: In his discussion of Objection 2 (which I have omitted here), Plantinga concedes that a believer in *austere* theism—the doctrine that there is an extremely knowledgeable and powerful being—escapes the epistemic difficulties besetting the naturalist only if he also believes something *in addition* to austere theism, namely, that we have been created in the image of this being. That paves the way for the third objection, which Plantinga epitomizes as "Why can't the naturalist just add a little something?" As Ginet has put the objection, "how is it that the theist is allowed to build into her metaphysical hypothesis something that entails R or a high probability of R but the naturalist isn't? Why isn't it just as reasonable for the naturalist to take it as one of

[50] Unless high intrinsic probability for R dictates otherwise, but that would also affect Plantinga's claim that R has inscrutable probability given N&E. We shall return to this point below.

[51] *WPF*, 233–234; I have suppressed four mentions of N&E that do not affect the present point.

[52] Perhaps someone will suggest that the revised Principle 1 does not let the rot begin, on the ground that the Super Bowl proposition derives its warrant from R itself. However, I can see no reason for saying this unless every proposition other than R derives its warrant from R itself, in which case the revised principle cannot be used for Plantinga's end of showing that R is defeated by N&E.

the tenets of naturalism that our cognitive systems are on the whole reliable (especially since it seems to be in our nature to have it as a basic belief)?"[53]

I trust that my discussion of Objection 1 will have lent additional urgency to the present objection. We saw in that discussion that given anything anyone believes, there is almost sure to be something else one believes that (by the standards of Principle 1) serves as a defeater for it. There will be no way of neutralizing the force of this defeater unless other things one believes on the side are allowed to do the job. That being so, why may the naturalist not have in his noetic structure—whether as part of his naturalism or not scarcely matters—beliefs that defeat defeaters of R?

Objection 4: The last objection I shall consider is the one Plantinga calls the "Maximal Warrant" or "Beyond Defeat" objection. He states it thus: "R has a great deal of intrinsic warrant for us. . . . It has so much intrinsic warrant, in fact, that it can't be defeated—or at any rate can't be defeated by the fact that P(R/N&E) is low or inscrutable."[54]

In reply to this objection, Plantinga concedes that R has intrinsic warrant. What he denies is simply that it has "a colossal degree of warrant, so much warrant, in fact, that it can't be defeated."[55]

That reply seems to me to misconstrue the thrust of the objection. The point of insisting that R has intrinsic warrant is not to set it up as an inherently indefeasible proposition. It is rather to authorize any of us—including the naturalist—to add it to his store of beliefs (or, more plausibly, to have had it there all along). And if the naturalist's noetic structure does contain R, he will not have a defeater for R in the mere fact that R has inscrutable probability given N&E. R will have high probability with respect to his total noetic structure even if it has inscrutable probability with respect to selected parts of it.

As I see it, then, the real function of Objection 4 is to shore up Objection 3: the intrinsic warrant of R may not make it inherently indefeasible, but it does make it quite all right (and not contrary to "true philosophy") for the naturalist to believe it.[56]

Let me summarize our discussion of Objections 1, 3, and 4. What underlies Plantinga's claim that the naturalist has a defeater for R is apparently Principle 1, which says that you have a defeater for one of your beliefs as soon as

[53] Carl Ginet, "Comments on Plantinga's Two-Volume Work on Warrant," *Philosophy and Phenomenological Research* 45 (1995): 407.

[54] "Naturalism Defeated," 15.

[55] Ibid., 52.

[56] There is another possible application of the point that R has high intrinsic warrant. According to Bayes' theorem, P(R/N&E) = P(N&E,R)P(R)/P(N&E). A high level of probability for R will therefore tend to increase the conditional probability of R on N&E. In fact, were P(R) very close to 1, we could assign P(R/N&E) a value lower than .5 only by supposing that R makes N&E less probable by nearly a half than it would be otherwise. So one possible point of relevance of the "high intrinsic warrant" idea is that it prevents R from having low probability on N&E in the first place. (I am using Bayes' theorem here in the opposite direction to Plantinga's use of it at *WPF*, 228.)

you believe something else on which it has low or inscrutable probability. But we have seen that Principle 1 (even in the qualified form we considered) is an epistemological land mine. It would imply that for anything any of us believes (including R), there is almost bound to be something else we believe that serves as a defeater for it (Objection 1). To avert the havoc Principle 1 threatens to cause for our knowledge, we must *either* reject it outright (in which case we do not have a basis for saying that N&E is even initially a defeater for R) *or* let the principle stand, but allow that the initial defeat relations it sets up may dissolve away when we take the rest of a subject's noetic structure into account. But in the latter case, why, indeed, can't the naturalist add to his noetic structure (or have there all along) something that would neutralize the defeating force of N&E—perhaps even R itself, as suggested in Objection 3? Why especially may he not do so if R has high intrinsic warrant for everyone, as suggested in Objection 4?

B. Is All of the Naturalist's Knowledge Destroyed?

Let us assume, contrary to what I have been arguing so far, that the naturalist does have a defeater (more accurately, an ultimately undefeated defeater) for R. Does it follow that all of his knowledge is thereby destroyed? The answer would be yes provided we also gave answers of yes to two further questions: (i) Does having a defeater for R entail having a defeater for everything? (ii) Does having a defeater for something entail not having knowledge of it?

Plantinga's answer to question (ii) is yes. He notes that a circumstance that defeats the *warrant* a belief has for me need not be a defeater in his sense, that is, need not make it *irrational* for me to continue to hold the belief.[57] But he evidently does not countenance the converse possibility, that a defeater might make a belief irrational for me while leaving its warrant (and thus its status as knowledge) intact. He writes, "If I acquire an undefeated defeater for a belief but continue to hold it, that belief will not have warrant for me."[58] That seems right. To maintain otherwise—to hold that warrant and knowledge can live through defeat—would be to allow that one may *know* things even though rationality requires one to withhold belief in them.

Plantinga's answer to question (i) is also yes, but here I am less sure that he is right. Why does having a defeater for R entail having a defeater for everything? If my belief in that one (admittedly very special) proposition becomes irrational, why do all my other beliefs become irrational, too?

[57] He cites the familiar example of the fake barns: oblivious to the fact that there are many fake barns in my neighborhood, I see a real barn and form the belief that I am seeing a real barn. In this situation, the presence of the fakes destroys the warrant my belief would otherwise have, but does not make my belief irrational. See "Naturalism Defeated," 23–24.

[58] "Naturalism Defeated," 23; see also *WPF*, 41.

One possible answer, conjectured by Sosa,[59] is that Plantinga subscribes to the *implicit premise thesis*—the thesis that whenever we accept any deliverance of a faculty, our doing so rests on the implicit premise that that faculty is reliable. That, of course, would explain why no belief is rational unless R is: R is an implicit premise behind any belief. It would also explain the difficulty Plantinga discerns in trying to bolster R by argument or defeat any defeaters of it: a premise enlisted in either of those causes could not be adopted rationally unless one was already rational in believing R. But how plausible is the implicit premise thesis? Sosa notes that that a child may be innocent of any belief in the reliability of her faculties, yet not on that account be irrational in her simple beliefs about food and shelter. I note as well that the implicit premise thesis seems quite out of keeping with the externalism that Plantinga defends at length in his two volumes on warrant. I note finally that it is also out of keeping with the externalism I earlier identified as crucial in escaping the Cartesian Circle.

Perhaps Plantinga does not hold that R is required as a positive premise behind any belief; perhaps he only subscribes to the negative condition that R not be something it would be irrational for the subject to believe. A point from DeRose is instructive here: "But, interestingly, even though a particular belief was formed with no thought about any relevant epistemic principle [assuring us of the reliability of the belief's source], the belief can be undermined and thereby be cast into doubt by casting doubt on the principle behind it."[60]

I am not convinced myself that even the negative condition is correct. Consider the position in which Descartes finds himself in the fourth paragraph of the Third Meditation. He confesses there that for all he knows, clear and distinct perceptions are not always true. As Plantinga might put it, the probability that clear and distinct perception is a reliable faculty is at best inscrutable for him, given his present knowledge of his maker. At the same time, he reports himself as being certain that various particular deliverances of that faculty are true—"Let him who will deceive me, He can never cause me to be nothing while I think that I am." Might this certainty not amount to *knowledge* on his part, even alongside his doubt about the general reliability of clear and distinct perception?

Perhaps the answer to this question depends in part on whether the doubt is a matter of its being positively unlikely that one's relevant faculties are reliable or simply a matter of one's simply not knowing that they are. In this connection, recall that Descartes characterizes his doubt about reliability as "very slight, and so to speak, metaphysical."[61] Recall also that Plantinga contends simply that R has either low *or inscrutable* probability given what the naturalist believes. Are those circumstances enough to keep Descartes or the natu-

[59] Ernest Sosa, "Plantinga's Evolutionary Meditations," chap. 5 in this volume.
[60] DeRose, "Descartes, Epistemic Principles," 226.
[61] HR 1:159.

ralist from knowing anything? If not, we should not maintain (as Plantinga does) both elements of the conjunctive proposition 'inscrutable probability of R in relation to something you believe gives you a defeater for R, and a defeater for R gives you a defeater for everything'.

Perhaps it is also relevant whether one's belief in the proposition that allegedly defeats R is itself warranted. Plantinga explicitly holds that a defeater becomes "operative" simply by my believing it, whether with any warrant or no. What if a philosopher comes to doubt the reliability of his faculties through bad arguments? Even if they are plausible (though unsound) arguments, mightn't a good design plan preserve his ordinary beliefs until he can find out what is wrong with the arguments? Sometimes I think skeptics who reason themselves into the suspicion that their faculties are unreliable know a great many things despite themselves.

C. Is the Theist in a Better Epistemic Position Than the Naturalist?

There is first of all the question whether theism really supports R. The relevant tenet of theism, Plantinga tells us, is that God has made us in his image. But that is evidently compatible with our falling drastically short of God in both power and goodness. Why may we not also fall short of God in knowledge—not just in the extent of it, but also in the reliability of it? On the other side of the coin, I wonder whether a New Age religionist who believes only in a "force for truth" might fare as well as an old-time religionist as far as Plantinga's argument is concerned.

Setting such questions aside for now, let us assume with Plantinga that theism provides positive support for R. How does that enable the theist to avoid the naturalist's predicament? Plantinga makes it clear that he envisions no argument up to the existence of God and back down to the reliability of our faculties, sensing in that procedure what he calls pragmatic circularity if not outright Cartesian circularity.[62] The situation is rather that the theist has nothing impelling him in the direction of skepticism to begin with—no element of his noetic system that serves as an initial defeater of belief in R.

But let us review what it was that made trouble for the naturalist. He was held to have a defeater for R in the mere fact that R's probability given N&E is low or inscrutable. In other words, he was said to have a defeater for R just because he is in the position of having to admit

> I believe N&E; and for all I know, R has low probability given N&E.

But if that were a bad position to be in, wouldn't the following position be just as bad?

[62] *WPF*, 234.

I do not believe N&E; but for all I know, (i) N&E is true, and (ii) R has low probability given N&E.

If the right attitude for one in the first position is, as Plantinga recommends, agnosticism toward R, wouldn't that also be the right attitude for one in the second position? If so, the naturalist's predicament is not avoided simply by *believing* theism in preference to it. One would have to *know* that the naturalist's worldview is wrong—that the troublesome N&E combination does not obtain.[63] That is to say, salvation in this case must be by knowledge rather than grace. And hasn't Plantinga made such knowledge difficult to come by?

Well, he has if he subscribes to the implicit premise thesis, but I expect to hear that he does not subscribe to that thesis. Moreover, in other places Plantinga has proposed a variety of roads to religious knowledge. Let us suppose that someone, perhaps by following one of these roads, has come to know that N&E is false and has never had occasion to doubt R. Such a one need not be perturbed by the considerations of the previous paragraph. It seems to me, however, that there are certain doubts that may come up and that will not easily be removed if Plantinga's strictures are correct. That means that the theist's position is potentially as precarious as that of Descartes's atheist geometer. Suppose, for example, that the theist reflects on the fact that his having been made in God's image is compatible with his falling far short of God in power and goodness, and is thus led to wonder whether he might fall equally short in the reliability of his cognitive faculties. Or suppose he considers those religious traditions in which human beings are primarily meant for spiritual goods that do not much depend on cognitive excellence. Might he then have doubts about R? The rot will have begun, and there will be no way of stopping it.

D. Conclusion

By way of conclusion, I would like to approach our issue from a different angle. Let us distinguish three positions on the relation of proposition R (= our faculties are reliable) to the rest of our knowledge. According to the first, R is *not* a proposition we need to know in order to know other things, since the satisfaction of certain external conditions (perhaps including R itself) suffices to give us knowledge. Call this the *externalist* position. According to the second, we *do* need to know R in order to know other things, and are fortunately enabled to do so by the fact that R is a self-evident first principle. Call this the *Reidian* position.[64] According to the third, R is again something we need to know in order to know other things, but it is not self-evident. We

[63] Strictly speaking, what one would have to know is that either N is false or E is false or R has adequate probability after all given N&E.

[64] I call the position Reidian because Reid lays it down as a self-evident first principle "that the natural faculties, by which we distinguish truth from error, are not fallacious." He goes on to say

know it (if at all) only by deriving it from other things we know. Call this (for obvious though contestable reasons) the *skeptical* position.

These positions are logically exhaustive; there may be variations within them, but not outside them.[65] If either the externalist or the Reidian position is correct, both theist and atheist may be in the business of knowing. If the skeptical position is correct, neither of them can be in that business. I find it hard to discern a position that makes knowledge possible for the theist but not for the atheist.

---

that "if any truth, can be said to be prior to all others in the order of nature, this seems to have the best claim; because in every instance of assent . . . the truth of our faculties is taken for granted, and is, as it were, one of the premises on which our assent is grounded." These quotations are taken from essay 6, chapter 5 of the *Essays on the Intellectual Powers of Man*, ed. B. Brody (Cambridge: MIT Press, 1969), 630–632. However, I do not think these passages represent Reid's best view. There is also a strong externalist streak in Reid, which I expound in "Reid on the First Principles of Contingent Truths," *Reid Studies* 3 (1999): 3–30.

[65] For example, Sosa takes up the first position with respect to animal knowledge and the third with respect to reflective knowledge—avoiding skepticism by allowing reflective knowledge to be generated by animal knowledge of R.

PART III

CONDITIONAL PROBABILITIES AND CONFIRMATION THEORY

A House Divided against Itself Cannot Stand

Plantinga on the Self-Defeat of Evolutionary Naturalism

Timothy O'Connor

Alvin Plantinga argues that belief in evolutionary naturalism is self-defeating. Let R denote the thesis that our basic cognitive faculties are mostly reliable, and EN the thesis that human beings and their cognitive faculties arose by means of entirely natural processes of the kind posited by current evolutionary biology, processes unguided and undesigned by God or any other supernatural being. The probability of R on EN, Plantinga plausibly maintains, is inscrutable by us. Since EN is relevant to the truth of R, the inscrutability of R on EN gives the adherent of EN a reason to withhold belief in R: EN is *evidence*, for the naturalist, that calls into question his belief in R. Withholding belief in R clearly would have disastrous implications for one's beliefs, as R's truth underpins the warrant for all our other beliefs. Worse still, Plantinga contends, there is no reasonable means of escaping this predicament once one is mired in it. Since the argument provides a defeater for all the naturalist's beliefs, he is left with nothing that might enable him to defeat the defeater. Hume's game of backgammon beckons.

Since R has a foundational status in our system of beliefs, we can evaluate the force of Plantinga's reasoning only in the context of a general account of epistemic justification or warrant. For the purposes of this essay, I will adopt Plantinga's own 'proper functionalism', on which "a belief has warrant, for a person, if it is produced by her cognitive faculties functioning properly in a

I thank Al Plantinga for helpful correspondence on the subject of this chapter in response to an earlier paper of mine ("An Evolutionary Argument against Naturalism?" *Canadian Journal of Philosophy* 24 [1994]: 527–540). During that time, many a breakfast of mine was given over to reflections such as those below—so much so, that for some time afterward, eating muesli regularly triggered thoughts concerning the machinations of evil demons. I have written the present essay partly in the hope that I may again eat my breakfast in peace.

congenial epistemic environment according to a design plan successfully aimed at [truth]."[1] As with other externalist theories, this account allows that cognitively basic beliefs such as R may be fully warranted despite the fact that a person cannot give a non-question-begging argument on its behalf. It simply need be *true* that our faculties function properly (and aim at truth) in giving rise to this belief—one needn't be in a position to argue that this is (or is likely to be) so.

This, at any rate, is how things initially stand for R, on Plantinga's account. His critique of belief in naturalism presupposes that as we begin to reflect on our beliefs concerning the origins of our cognitive faculties and their probabilistic connections to R, we should begin modifying our confidence in R. In the language of probabilistic reasoning, it is appropriate to 'start off' assigning a very high probability to R, independent of any opinion we may have concerning the absolute a priori probability of R. Subsequently, however, we should *conditionalize* the value we assign to R on factors that are relevant to the probability that R is true. If we believe such factors to obtain and judge the probability of R on them to be inscrutable, we should declare all bets off regarding the truth of R and suspend belief. We should do so because this is precisely how we should proceed in analogous circumstances with more mundane beliefs. Suppose I am driving through the countryside and come to believe there is a sheep in the field. But then I am told by a humorless and reliable friend that this is yet another locale in which the residents like to play tricks on strangers by carefully dressing up dogs to look like sheep. Since I don't know the proportion of sheep to look-alikes in this vicinity, the reasonable course is to refrain from believing that what I see is a real sheep.

Are matters as straightforward as this with respect to R? Consider two examples that Plantinga himself discusses in other contexts. The first is the Case of the Purloined Letter. You are presented with powerful evidence that you have stolen an important letter. You had motive, means, and opportunity. Sadly, it is known that you are not above such shameful activities as filching letters. Finally, there is very strong evidence that you were the only person who could have stolen the letter and it was later recovered from your bureau drawer. Now, before hearing the prosecutor's case, you believed that you had not stolen the letter. And with good reason: you had no reason to doubt your memory in this regard (which was in fact functioning normally), and you distinctly recalled seeing the letter and resisting the temptation to take it. Should you change your belief in the face of the powerful evidence to the contrary? Presumably not. You have countervailing evidence in the form of a distinct memory of not having taken the letter.

Second case. A Christian theist reflects on the dismal facts of pain and suffering in human history. He knows of no convincing account of adequate reasons God might have for permitting these things to occur. He then considers

[1] *Warrant and Proper Function* (New York: Oxford University Press, 1993) 237.

the probability that God exists on this evidence. Suppose that, impressed by our finitude, he judges it unlikely that we could figure out what God's reasons would be for creating a world with such features, were He to exist and have created a world just like ours. He concludes that the conditional probability that God exists on facts about the magnitude and distribution of human suffering is inscrutable. Should he become agnostic? As Plantinga himself emphasizes, if his proper functionalist account of warrant is basically correct, we cannot divorce this question from others concerning God's existence and His intentions concerning the functioning of human cognitive capacities. It is a plausible to suppose that if traditional Christian teaching is essentially correct, then God does not intend for human beings to cease believing in Him under such circumstances. On such teaching, a Christian who confidently persists in belief despite not knowing what to say about the evidential bearing of the facts of evil on God's existence is, other things being equal, regulating his beliefs in accordance with God's design plan. It is the tendency toward doubt or disbelief that reflects noetic malfunction.[2]

These examples highlight the fact that simple conditionalization on new evidence need not always be required within a proper functionalist epistemology. Some beliefs will have a special status, depending on certain facts about the *design plan* of the cognizer in question. An evolutionary naturalist will naturally begin his reflections on Plantinga's challenge by considering the implications EN itself might have for thinking about proper function, just as the theist does in considering the challenge to theism from evil. An astute naturalist will note first of all that the argument doesn't, as a psychological matter, tend to push people to cognitive despair, as Plantinga claims it ought to do. Our belief in R is quite tenacious. Given EN, this is not surprising. There is no plausible evolutionary story one might tell about why our cognitive faculties should be designed to *regulate* this belief strictly in accordance with our evidence for it. Doing so would obviously be disastrous for our coming to true beliefs about other matters. But it would equally undercut our ability to come to a true belief about R itself. It is plausible, given EN, that the naturalist's unshaken confidence in R reflects our cognitive design plan, stemming from a module aimed at truth. We're designed to start off accepting R without evidence and to continue on that way. In view of our cognitive limitations—we're not able to pull off any version of Descartes's project—it is proper that we adopt R and related beliefs as unquestioned framework principles, against which we can sensibly adjudicate our beliefs about less foundational matters.

I think we can say something even stronger. Strictly speaking, there *couldn't* be a defeater for R, for any creatures in any possible world. That would re-

[2] More needs to be said about the instantiation of a proper functionalist epistemology within a Christian theistic metaphysics to make these claims convincing, of course. There is no better place to begin exploring this matter than Plantinga's own *Warranted Christian Belief* (New York: Oxford University Press, 2000).

quire a *design plan* which reliably aimed at truth in some circumstances, a part of which was that we ought to give up belief in R when we take note of certain of our beliefs about the world and about related conditional probabilities. But as that potential upshot would lead one to abandon the attempt to form true beliefs, it surely could not be part of a design plan with that very end. Or if the concept of a design plan is to be quite broad, then only egregiously flawed instances would contain in this way internal obstacles to their own success.

At this point, Plantinga might protest that our argument has proven too much. Are we not claiming that R is necessarily immune from evidence? And isn't this plainly wrong—couldn't one have, for instance, good reason to think one was or was in the process of becoming insane?

But we are not claiming that R is indefeasible. One could *rationally* conclude that R is false (in relation to oneself) by having powerful, direct evidence that there *is no design plan successfully aimed at truth*. This would seem reasonable if one's cognitive output were persistently and massively inconsistent—which, I take it, might well capture the circumstances under which one might rationally take oneself to be insane.[3] Of course, once one draws this conclusion, it will undercut one's confidence in the very evidence that led to it, thereby giving rise to Hume's loop of reason against itself, at least until one retires to more soothing pursuits such as backgammon. But note that we shouldn't say that one would be *warranted* in drawing this fateful conclusion, for that term of art is tied to the notion of a design plan aimed at truth, and there is no purchase on the idea of such a plan that generates persistent, massive inconsistency. (What then is the operative notion of 'rational' here? It's hard to say precisely—perhaps a basic kind of theoretical rationality which requires a minimal degree of evidential consistency.)

Plantinga, however, argues that massive internal inconsistency is not the only way one might reasonably be led to abandon R. For suppose I come to believe that I am the victim of Descartes's evil demon, who is bent on my being deceived about my true situation. Don't I then have good reason to abandon R?[4] To evaluate this analogy properly, we must first tighten it up a bit. As the evil demon hypothesis is usually presented, we know that if it's true, then R is false. (By hypothesis, all of my beliefs about the material world are false.) But in the challenge posed to the naturalist, we are urged to say only that the probability of R on evolutionary naturalism is inscrutable. So let us modify the evil demon story accordingly. We will now say that I believe that there is an evil demon who has created my cognitive faculties, but I judge

[3] Perhaps a second way is this: rather than believing merely that the general probability of R on a certain broad *type* of origin for our cognitive faculties is low or inscrutable, we come to believe confidently that the specific circumstances of the actual past, together with the evolutionary processes at work, were such that R is improbable on them. But how could we know that, apart from the kind of direct evidence (pervasively inconsistent outputs and the like) discussed in the text?

[4] "Naturalism Defeated" (unpublished), 51–52.

the probability of R on this origin to be inscrutable. (Perhaps I believe that this demon has created numerous races on various planets, and in at least one case has chosen to massively deceive them, and in at least one other case has chosen to give them reliable faculties. But I have no idea whether either case was an isolated act of whimsy or reflected a decided tendency on his part.)

Once we make this change in our hypothesis, however, it isn't plausible that we are rationally required to abandon R. There isn't massive inconsistency of the sort noted just above, a sort that would make it unreasonable to continue believing there is any coherent design plan underlying one's belief formations. And whether R would have or lack warrant under those circumstances depends on the demon's design plan: has he so designed the cognitive system in question that a properly functioning specimen would abandon R under such circumstances? Well, if the agent is like most naturalists who encounter Plantinga's argument, he is not *inclined* to do so. So the only evidence the agent has in this matter—his own tenacious belief—suggests that he is functioning properly in continuing to believe R. Nothing in the structure of Plantinga's argument as applied to this case would give him reason to think otherwise.

Let me bring the issue into sharper focus by yoking the cases of the Christian theist's response to the evidential bearing of evil on theism and the naturalist's response to the evidential bearing of evolutionary naturalism on R, and then posing a dilemma. Plantinga would say that the probability of theism given the facts of pain and suffering as we know them—P(T/E)—is (at best) inscrutable for the theist. Must the theist, if he is to adjust his beliefs properly, conditionalize on his ignorance, so to speak? There are four possibilities here:

(1) He must conditionalize on his ignorance and thereby come to regard the probability of theism itself as inscrutable for him in the circumstances. If one accepts this, one is also likely to accept Plantinga's own verdict on the case of naturalism. But the larger upshot will be that *every* reflective person should come to withhold belief in R!

(2) He must conditionalize on his ignorance, but he need not come to judge the probability of theism as inscrutable. For he has an independent source of warrant that neutralizes the potential defeater of his belief before it takes hold: the internal testimony of the Holy Spirit that sustains the Christian's confident trust in the existence and character of God. This is a kind of evidence, in a broad sense of the term, that counteracts the potential impact of the inscrutability of P(T/E) for him.[5] But if the theist should say this, why should not the naturalist say the same concerning the 'evidence' coming from his own tenacious belief that R?

(3) He need not conditionalize on his ignorance or modify his beliefs in

[5] A brief remark by Plantinga in correspondence several years ago suggested this interpretation to me. But he appears to reject this view of the matter in *Warranted Christian Belief*, 478ff.

any other way. Theism continues to have warrant for him. (But then by parity of reasoning the same should hold for the naturalist.)

(4) He need not conditionalize on his ignorance, but he should adjust his belief concerning P(T/E). If theism is true, it is part of my design plan that I should believe T despite my inability to discern a priori P(T/E). But since it is also part of my design plan that I adjust my beliefs, as best I can, in the direction of greater probabilistic consistency, I should *conclude* that P(T/E) is high. Put differently, I should see on reflection that T's having a good deal of warrant for me commits me to believing tacitly that the probability of T on all the relevant evidence I possess is high. And because of the special status T has in my God-given design plan, it should trump any beliefs I may hazard concerning conditional probabilities governing T, i.e., my belief in T has more warrant than do any of my beliefs concerning the conditional probability of T on any particular proposition. Although Bayesian accounts of theory confirmation have a difficult time handling this fact, we clearly can rationally revise our estimates of conditional probabilities. (Consider a scientist who comes better to understand the implications of a complex theory and so revises his estimates of the probability of that theory on various bits of potential evidence.) But again, if this is so for the theist, what reason have we to assume it is different for the naturalist? If one is to challenge the propriety of the naturalist's doing this, it cannot depend on assuming facts about the design plan inconsistent with naturalism.

Of these four options, (1) will be implausible to all nonskeptics, and (2) strikes me as stretching the notion of evidence beyond useful limits, so that it becomes little more than a terminological variant on (4). Whichever way we go on (3) or (4)—a verdict that will reflect one's judgment on whether it is appropriate to extend the machinery of probability theory to belief kinematics quite generally—since parallel moves are ready to hand, the naturalist has nothing to fear from Plantinga's argument.

Conditional Probabilities in Plantinga's Argument

Richard Otte

Alvin Plantinga has published an unusual argument that purports to show that naturalism is irrational. This argument relies upon the conditional probability that our cognitive faculties are reliable (R) given naturalism and evolutionary theory (N&E). If that probability is low, Plantinga has a preliminary argument that theism is to be preferred to naturalism. If the probability is either low or inscrutable, Plantinga's main argument claims that the naturalist is self-referentially incoherent. Thus the naturalist has a defeater for every belief he has, including naturalism, and his noetic framework is irrational.

One of the most basic objections to these arguments is that they make use of the wrong probabilities. It is well known that probability is very sensitive to the information conditionalized on. Plantinga's arguments look at the probability of R conditional on N&E, but this ignores other information we have that would make R likely. In effect this objection claims that Plantinga's argument ignores relevant evidence we have for R, and thus it is no surprise that P(R/N&E) is low. But when we take into consideration this additional relevant evidence, we find that the probability of R conditional on it is high. If this objection is correct, Plantinga's arguments will not even get off the ground.

In this chapter I will investigate in detail objections of this sort. I will begin by looking at this objection applied to Plantinga's preliminary argument. This objection has a natural response within the preliminary argument. Although this response is successful, it does not answer most of the important questions about conditionalization in this argument, and the nature of the preliminary argument is such that it is not as forceful or convincing as the main argument. Thus I will focus most of my attention on Plantinga's main

argument. The response to this objection given within the preliminary argument is not available to defend the main argument, and to answer this objection a deeper inquiry into what should be conditionalized on is required. This will allow us to propose a more satisfactory solution to the problem of what to conditionalize on in Plantinga's arguments. The solution to this problem has consequences for Plantinga's claim that P(R/N&E) is low or inscrutable, and we will then give a new argument for this by taking into account my proposal of what to conditionalize on.

Plantinga's Preliminary Argument

Plantinga's preliminary argument considers P(R/N&E) upon different views about the connection between our behavior and the content of our beliefs. On most of these views, such as the view that there is no causal connection between them, we have no reason to think that R is likely. But Plantinga thinks the probability of R is not high even on the view that the content of our beliefs can causally affect our behavior. The main reason for this is that any given action can be produced by many different belief-desire combinations. Plantinga gives the example of Paul, a prehistoric hominid in a situation where he sees a tiger and needs to display tiger-avoidance behavior. As Plantinga puts it:

> But clearly this avoidance behavior could be a result of a thousand other belief-desire combinations: indefinitely many other belief-desire systems fit B equally well. . . . Perhaps Paul very much *likes* the idea of being eaten, but whenever he sees a tiger, always runs off looking for a better prospect, because he thinks it unlikely that the tiger he sees will eat him. This will get his body parts in the right place so far as survival is concerned, without involving much by way of true belief. . . . Or perhaps he thinks the tiger is a large, friendly, cuddly pussycat and wants to pet it; but he also believes that the best way to pet it is to run away from it. . . . or perhaps he thinks he is about to take part in a sixteen-hundred-meter race, wants to win, and believes the appearance of the tiger is the starting signal; or perhaps. . . .[1]

In response to examples such as these, some deny that our beliefs could be mostly false and adaptive. Plantinga responds with an example of a person who believes everything is a witch and all his ways of referring to things involve definite descriptions entailing witchhood. This person's beliefs would be adaptive but unreliable. Since our beliefs could be adaptive and mostly false, and since natural selection only selects for behavior and not the con-

[1] *Warrant and Proper Function* (New York: Oxford University Press, 1993), 225–226.

tent of beliefs, Plantinga claims that the probability of R will not be high even if the content of our beliefs causally affects our behavior. Since P(R/N&E) is a weighted average of positions describing the causal connection between the content of our beliefs and behavior, this probability will not be high.

Plantinga then claims R is very likely on the assumption of theism (T). According to traditional theism God created humans in his image. Part of the image of God is having knowledge, and thus we are created with reliable cognitive faculties. P(R/T) is high because conditional on theism we have no reason to doubt our faculties are reliable.

Plantinga then uses Bayes' theorem to compare the support theism and naturalism receive from our cognitive faculties being reliable.

$$\frac{P(T/R)}{P(N/R)} = \frac{P(R/T)\,P(T)}{P(R/N)\,P(N)}$$

If this ratio is greater than one, then theism receives more support from R than naturalism does. If it is less than one, then conditional on R naturalism is more likely than theism. Since P(R/T) is much greater than P(R/N), Plantinga says that we have an initial prima facie reason to prefer theism over naturalism.

Conditionalizing on Other Information

A major response points out that by conditionalizing only on R in this argument we ignore relevant evidence we have. Although P(R/N&E) may be low, when we conditionalize on other things we know (O), we find that P(R/N&E&O) is high. For example, many naturalists don't take seriously the position that the content of our beliefs does not causally affect our behavior; they ignore this possibility. The idea is that many naturalists are certain that the content of their beliefs affects their behavior, and thus we should conditionalize on that. This makes R more likely than when the other options are factored in. Other relevant things we believe are that we live in society and can communicate with others; we are able to carry on complicated conversations. Furthermore, we have been enormously successful in developing successful scientific theories. Many of these things are uncontroversial, and the idea is that they would be very unlikely, if not impossible, unless our cognitive faculties were generally reliable. It is claimed that if we conditionalize on these factors we find that the probability of our faculties being reliable is high.

Although this is a common response to Plantinga's argument, I am not convinced that these other factors mentioned actually raise the probability of our faculties being reliable. For example, I see no reason to think that we would have difficulty living in society and communicating if our beliefs were mostly false; it seems quite possible that society could get along just fine with shared false beliefs. Plantinga's witch example illustrates this nicely; a society that believed and referred to everything as a witch would be able to communicate and carry on very complicated conversations even though their beliefs were generally false. Furthermore, the argument that successful science requires true theories has received extensive discussion in the literature on scientific realism. There are good reasons for rejecting the idea that successful science requires true or approximately true theories, and there is no convincing argument that shows that theories are likely to be true if they are successful. However, for the sake of argument we can ignore these problems and assume that there are additional factors that make it likely that our faculties are reliable. It will turn out that Plantinga's preliminary argument is not weakened by conditionalizing on these factors.

To assess this response we need to look at exactly what probabilities are involved in Plantinga's argument. Plantinga argued that R supports T more than it supports N; P(N/R) is lower than P(T/R). He used Bayes' theorem to show P(N/R) and P(T/R) depends on P(R/N) and P(R/T). We have seen that P(R/N) and P(R/T) ignore relevant information we have, and so we need to conditionalize on additional information. But we cannot simply substitute P(R/N&O) for P(R/N) into the above form of Bayes' theorem; the result would not be a theorem of probability.

Fortunately we can construe the objection as proposing that instead of looking at P(N/R), we look at P(N/R&O) instead. Instead of comparing P(N/R) with P(T/R), we need to compare the probabilities of naturalism and theism conditional on all of this additional information. There is a lot to be said for this proposal. One of the least controversial requirements of confirmation theory is the requirement of total evidence, and this proposal can be seen as simply saying that in assessing the probabilities of naturalism and theism we need to look at all the relevant evidence, not simply the evidence that our faculties are reliable. We can then use a version of Bayes' theorem to get:

$$\frac{P\,(T/R\&O)}{P\,(N/R\&O)} = \frac{P\,(R/T\&O)\,P(O/T)\,P\,(T)}{P\,(R/N\&O)\,P(O/N)\,P\,(N)}$$

This version of Bayes' theorem has more terms than the original, due to the additional information being conditionalized on. We are assuming that the new information conditionalized on raises the probability of R given N, P(R/N&O) > P(R/N). But this does not imply that P(N/R&O) > P(N/R), nor

does it imply that naturalism is more likely than theism given this information. If we conditionalize on some other information we need to also look at how probable this other information is, conditional on N and on T; that is why the above formula also contains the terms P(O/N) and P(O/T). We can conditionalize on some other information O, but we then have to look at how likely that information O is, given naturalism and given theism. This strategy of conditionalizing on some other information O that raises the probability of R does not guarantee that N will be more probable on O and R than N was on R, even if R is more probable on N and O. Even if this other information raises the probability of R given naturalism, this other information may be improbable given N. In that case this will not be a successful response to Plantinga's argument, because P(N/R&O) may still be much less than P(T/R&O).

To illustrate how this works consider conditionalizing on some of the other information naturalists have proposed. Suppose the additional information O is that the content of our beliefs causally affect our behavior. It then appears true that P(R/N&O) is greater than P(R/N). But the problem is that we have no reason to think that P(O/N) is high; that is, we have no reason to believe it is likely that the content of our beliefs causally affects our behavior given naturalism. Plantinga has argued that this O is actually very unlikely on N, and thus his preliminary argument is not threatened by using this new form of Bayes' theorem to conditionalize on O. We need to multiply P(R/N&O) by P(O/N) in the above formula, and so even if P(R/N&O) is high, their product will be low if P(O/N) is low. Without a convincing argument that this additional information is likely given naturalism, conditionalizing on this additional information does not weaken Plantinga's argument. Furthermore, the naturalist has given no reason to think that P(R/N&O)P(O/N) > P(R/T&O)P(O/T).

As another example suppose the naturalist were to claim that the additional information O is that we have won the cosmic lottery; the initial conditions were just right to allow us to develop and have mostly true beliefs. It is then true that P(R/N&O) and P(R/T&O) are quite high, simply because P(R/O) is high and neither naturalism or theism defeats that. But this also does not help the naturalist because we also need to look at P(O/N) and P(O/T): how likely is it that we won the cosmic lottery conditional on naturalism and on theism? It appears that P(O/N) is quite low; on naturalism there is no reason to think we won the cosmic lottery. Plantinga's original arguments that show that P(R/N) is low apply equally well to show that P(O/N) is low. Furthermore, there is reason to think P(O/T) is high. On the assumption of theism it is very likely that the initial conditions would be such that we would end up with mostly true beliefs. According to theism God wants us to have mostly true beliefs, and thus he would set up the initial conditions to bring this about. Taking account of this we see that this information O actually supports

theism more than it supports naturalism.[2] Thus even conditionalizing on the information that we won the cosmic lottery does not help naturalism in this preliminary argument. The same is true of all the other information proposed to conditionalize on.

We thus see that conditionalizing on additional information is not a good response to Plantinga's preliminary argument. Admittedly, some additional information will raise the probability of R for the naturalist. But the problem is that the additional information we considered turns out to be less probable on naturalism than on theism. Thus this strategy of replying to Plantinga's preliminary argument fails. Indeed, much of the additional information proposed to conditionalize on supports theism more than naturalism.

Of course, there are several things it would be possible to conditionalize on that would change the probabilities in favor of naturalism over theism. For example, suppose we conditionalize on something such as God does not exist or anything else that is a consequence of naturalism. Although these factors do change the probabilities in favor of naturalism, it is clear that it is improper to conditionalize on them. The problem of what can be conditionalized on in comparing two theories is difficult, although it is obvious that certain things should be excluded. We will give a more detailed treatment of

[2] In this preliminary argument we are interested in whether the evidence R&O supports naturalism more than theism, but there are different types of support relations. Plantinga seems to be working with an absolute sense of confirmation, where e confirms h if and only if $P(h / e) > .5$. However, another important support relation is incremental support: e confirms h if and only if $P(h / e) > P(h)$. If we look at incremental confirmation we find that R&O confirms T more than it confirms N. One way to judge the effect of R&O on N and T is to compare the following ratios:

A. $\frac{P(T/R\&O)}{P(N/R\&O)}$

and

B. $\frac{P(T)}{P(N)}$

If ratio A is greater than ratio B, then the evidence R&O supports theism over naturalism. If ratio A is less than ratio B, then R&O supports naturalism more than it supports theism. It will be easier to compare these ratios if we replace ratio A with its equivalent ratio A':

A'. $\left[\frac{P(R/T\&O)}{P(R/N\&O)}\,\frac{P(O/T)}{P(O/N)}\right]\left[\frac{P(T)}{P(N)}\right]$

Looking at A' it is easy to see that whether it is greater or less than ratio B depends completely on the ratio in the left bracket. Thus in comparing ratios A and B we can focus our attention on the following ratio:

C. $\frac{P(R/T\&O)}{P(R/N\&O)}\,\frac{P(O/T)}{P(O/N)}$

If this ratio is greater than 1, then the evidence R&O incrementally confirms theism more than it supports naturalism, since ratio A is greater than ratio B. If O is that we won the cosmic lottery, we have $P(O / T) > P(O / N)$ and $P(R / T\&O) = P(R / N\&O)$. From this it follows that R&O incrementally confirms theism more than naturalism.

what should be conditionalized on in our discussion of Plantinga's main argument below.

Throughout our discussion of this preliminary argument I have ignored the prior probabilities of naturalism and theism that occur in the above formulae. One of the lessons learned from Bayesian confirmation theory is that any adequate theory of confirmation must take account of the prior probabilities. Looking at the above formulae, we see that if the prior probability of naturalism is much higher than the prior probability of theism, this can offset the advantage theism receives from other parts of the formulae. If the naturalist believes the prior probability of theism is sufficiently low, then Plantinga's preliminary argument will not give him a good reason to prefer theism over naturalism. However, the naturalist might assign prior probabilities in such a way that Plantinga's argument does point to a problem, and his posterior probability of theism may be higher than his posterior probability of naturalism. This will depend upon the estimates of the prior probabilities.

We see that Plantinga's preliminary argument is partially successful and that one of the major objections to it is unsuccessful. The major objection raised against it does not weaken the argument. But with Plantinga, we should also realize that his preliminary argument only provides one reason to prefer theism over naturalism. There may be other reasons to prefer naturalism over theism. For example, the prior probabilities might be such that the person should remain a naturalist or there may be some other evidence against theism.[3] There are many other issues connected with this preliminary argument, but we now turn to Plantinga's main argument.

Plantinga's Main Argument

Because of the inherent weaknesses in the preliminary argument, Plantinga's main argument focuses on epistemic incoherence within naturalism instead of comparing naturalism with theism. In the main argument Plantinga assumes that P(R/N) is either low or inscrutable. From this Plantinga concludes the naturalist has a defeater for all of his beliefs, including N, and thus the naturalist is self-referentially incoherent.

At this point the naturalist may again object that Plantinga's argument makes the mistake of ignoring evidence that we have for R. Although P(R/N&E) may be low or inscrutable, we have other evidence O such that P(R/N&E&O) is quite high. This strategy of conditionalizing on additional

[3] Even if there is other evidence against theism, individual pieces of evidence can combine in unexpected ways. For example, one could have two pieces of evidence e_1 and e_2 that individually lowered the probability of naturalism, but together they might raise the probability of naturalism. How individual pieces of evidence relate to belief is not simple. This is a further problem with the preliminary argument.

information was not helpful to Plantinga's preliminary argument, because we then had to look at how likely that information was conditional on N and on T. But in the main argument Plantinga does not use Bayes' theorem to compare P(N/R) and P(T/R), and instead he focuses directly on P(R/N&E). In his main argument we can conditionalize on additional information without having to factor in how likely that information is, given naturalism. So the strategy of conditionalizing on additional information is much more promising in the main argument than it was in the preliminary argument.

A good way to think of this is to look at conditionalizing on additional information as a way of narrowing down the worlds that we are interested in. We begin by looking at how likely R is, given that we could be in any possible world. We then conditionalized on N&E, which restricts the worlds being considered to those in which N&E is true. We can progressively narrow down the class of worlds further, by considering only worlds in which there are beings, worlds in which there are beings with beliefs, worlds in which the content of these beliefs causally affects the beings' behavior, worlds in which these beings live in society and communicate, worlds in which these beings develop advanced science, worlds in which massive mistakes are not made, worlds in which these beings have mostly true beliefs, etc. It may be that when we look at the larger set of worlds, R is unlikely; but when we narrow down the set of worlds, R becomes more and more likely until it is certain. The objective probability of R given these sets of worlds exists for each set; the problem is not to determine which of these probabilities exist. The problem is one of deciding which probability to use in Plantinga's argument. Obviously it would not be useful to conditionalize on the sets of worlds at the very end of this progression; that would beg the question by conditionalizing on R itself. Nor would it be proper to conditionalize on only N&E; we know that certain worlds need to be excluded in determining the probability of R. So we need to find a way to limit the worlds we are interested in.

Solving the general problem of what to conditionalize on in cases like this is very difficult; I know of no good answer to it. But fortunately we do not have to solve the general problem in order to have an acceptable proposal for Plantinga's argument. By looking at Plantinga's argument we can discern some principles to guide us, without giving a complete solution to this problem.

Plantinga's argument *uses* conditional probabilities to argue for a philosophical conclusion. Thus the issue is one that deals with the application of probability: what probability statement should be used in this argument. In this case, Plantinga is using probability in the context of looking at how likely it is that our beliefs are reliable, given that naturalism and evolutionary theory is a correct account of the source or origin of our beliefs. Therefore we must certainly conditionalize on naturalism and evolutionary theory. In addition, since we are looking at the internal rationality of being a naturalist, what we conditionalize on must be limited to information accessible to us. A fur-

ther restriction on what to conditionalize on appears when we look at this discussion as a debate over whether a certain source or origin of our beliefs or faculties would be reliable. Because of this we cannot conditionalize on anything that presupposes or is based upon that source being reliable. Such a probability statement would not be useful in this discussion, because it would presuppose what is being questioned. For example, in Plantinga's argument if we were to conditionalize on R itself or that we won the cosmic lottery, then we could come up with a true probability statement, but not one that was useful. Doing so would beg the question in an important sense, because it assumes the source in question is reliable. On the other hand, we must also conditionalize on all information we have that is independent of or not epistemically based on the source in question. If we were to ignore any of this evidence, then the probability would be useless, since the ignored evidence may affect the probability. So we see that we have to conditionalize on all information we have that is independent of or is not epistemically based on the source of belief in question. I realize this is not very precise, but I hope it will give an intuitive idea of what should be conditionalized on in arguments of this sort.

Applying these ideas to Plantinga's argument, we can view his argument as raising the serious possibility that if naturalism and evolutionary theory are the source of our cognitive faculties, then all of the beliefs that we form are unreliable. If a source of some of our beliefs is unreliable, all beliefs proceeding from that source are suspect. Since Plantinga claims that naturalism and evolutionary theory is an unreliable source of our cognitive faculties, *all* of our beliefs are called into question. Thus we would be begging the question against his proposal if we were to conditionalize on the truth of any of the beliefs that arise from this source. But that means we cannot conditionalize on the truth of any of our beliefs, since they all arise from this source. What then is left to conditionalize on? Our experience. In Plantinga's argument we should look at the worlds that are consistent with our experience. But since one can only conditionalize on propositions, we should conditionalize on a proposition describing the experiences we have had, which we will denote by 'EXP'.[4] By experience I mean to include both what we might call 'external' experience, which would be our experience of the external world, and 'internal' experience, which would be experience of what we believe, desire, remember, etc. Although it is very difficult to characterize what experience is, I here use the term in a way that does not presuppose that what we experience is veridical. Thus we could experience seeing a tree, even if there were no tree there. Although there are many well-known problems surrounding this no-

[4] I do not wish to assume that the epistemic support a belief receives from an experience is the same as the support it receives from a proposition describing that experience. In fact, I think these can quite often be different. This difference points to the need for a theory of probability in which one can conditionalize on experiences as well as propositions. But this is not the place to develop such a theory.

tion, they do not affect this proposal. Our experience conditionalized on would include the following: we experience that we believe that our beliefs causally affect our actions, that we live in society, that we can communicate with each other, that we do not make massive mistakes, that we tend to correct the mistaken beliefs we do form, that our science is quite good, and that our cognitive faculties are reliable. This information conditionalized on will not include propositions such as our beliefs causally affect our actions, we do not make massive mistakes, our science is getting closer and closer to the truth, or that our faculties are usually reliable; but we do conditionalize on the information that we believe those propositions.

In proposing to conditionalize on EXP we are not assuming that the agent has any beliefs about his experience, nor do we assume that his introspective beliefs are reliable. We are discussing what probability should be *used* in reasoning about whether the naturalist is internally incoherent. If we are investigating the reliability of another person's cognitive faculties it is clear that we should conditionalize on their experience and not on the truth of their beliefs. If they happen to have some true beliefs about their experience, we still conditionalize on their experience; the information is conditionalized on because it is their experience, not because it is something they believe to be true. This is unproblematic in reasoning about others, but when we investigate the reliability of our own cognitive faculties a pragmatic problem arises. In working through Plantinga's argument the naturalist must apply the argument to his own case, and in order to do that he must have some beliefs about what his experience and beliefs are.[5] But although some beliefs are required in order to reason or think, these beliefs can be defeated. They are not immune to defeat simply because they are necessary in a pragmatic sense. If Plantinga is correct, these beliefs are defeated, and there is no epistemic ground for holding them. These pragmatic difficulties arise in the context of reasoning about the reliability of one's faculties; they do not point to any problem in it being correct to conditionalize on EXP.

To support my proposal, it will be helpful to look at what it is proper to conditionalize on in a few analogous cases. In these cases it will turn out that my proposal gives the intuitively correct answer of what to conditionalize on. Specifically, in each of these cases we will see that it is improper to conditionalize on beliefs produced by the source in question. Let us first consider an alternative hypothesis about the source of our perceptual beliefs that Plantinga often appeals to: the brain-in-a-vat scenario. Suppose that we are brains in vats and all our external experiences are given to us by the scientists maintaining the vats. Our perceptual beliefs will then be completely unreliable, although the scientists do not manipulate our other faculties. Applying Plantinga's reasoning to this, we are interested in the probability that our perceptual

[5] Of course, the agent need not have precise beliefs about what his beliefs are; he might have opinions about how likely it is that certain beliefs are his beliefs.

faculties are reliable, given we are brains in vats. Intuitively this probability should be low, if we conditionalize on the right information. What is called into question in this situation is the reliability of our perceptual beliefs; our other beliefs based on other faculties are not being questioned in this situation. Thus we can conditionalize on beliefs produced by other faculties, such as our a priori beliefs, since their reliability is not at issue. But we cannot conditionalize on the truth of any of our perceptual beliefs; if we were to do so, we would get the wrong result that our perceptual beliefs are highly reliable given that we are brains in vats. We should conditionalize on having these perceptual beliefs, since that is given by our internal experience, but we should not conditionalize on the beliefs being true. We see that in the brain-in-a-vat scenario we should not conditionalize on the truth of any beliefs that are based on the faculty in question. This intuitively correct answer is the same as that arrived at by the reasons given by my proposal.

As a variant of this example suppose you believe that you are now a brain in a vat, and that before placing you in the vat the scientists told you they occasionally program some brains with mostly correct perceptual experiences. These can be identified because of a certain red tint that accompanies their experiences. If someone has this belief and notices that their experience has that red tint, then they would have reason to think that their perceptual faculties are generally reliable. To ignore this red tint and the information it provides would be a mistake, since neither the information nor experience is based on the faculty in question. This shows the importance of conditionalizing on all information that is not rendered suspect by the source in question, as my proposal dictates.

Now consider an example where the only source of our beliefs about something is due to a certain instrument. Let us go back in time to when optical microscopes were our only source of information about certain microscopic phenomena; various other types of microscopes had not been invented. At this time nobody really knew how microscopes worked, and there was serious debate about whether microscopes were reliable instruments. As Ian Hacking has noted, well-known and respected scientists refused to allow them into their labs, because their reliability was unknown.[6] There was serious debate about whether what was seen in a microscope was real or merely an artifact of the instrument (there were cases of each, as it turned out). In this situation there was no good reason to believe what was seen in the microscope was real, since it was not known if the microscopes (the origin of the beliefs in question) were reliable. Thus intuitively the probability of beliefs about microscopic objects based on these microscopes was low. Now consider what we should conditionalize on in determining how likely it is that our beliefs based on the microscope were reliable. We should certainly conditionalize on our

[6] Ian Hacking, "Do We See through a Microscope," *Pacific Philosophical Quarterly* 62 (1995): 305–322.

general scientific knowledge of the time, since that is not based on the microscope. We should also conditionalize on our experience of what was seen in the microscope. But it would be improper to conditionalize on the truth of beliefs formed about the microscopic world as a result of the microscope. If we did, the probability that the microscope is reliable would be high, even though there are good reasons to doubt its reliability. For example, suppose we have a poor microscope that produces very unreliable images; most of the information it gives us is merely an artifact of the microscope and does not correspond to any properties of the thing being observed. If we were to conditionalize on the truth of beliefs formed by looking through this microscope, we would find that our beliefs matched what was known to be true, and thus it would be very probable that this unreliable microscope is reliable. Since the microscope is unreliable, to avoid concluding it is reliable we must not conditionalize on any beliefs whose source is in question, which in this case is the microscope. Thus we see that in this example the intuitive answer also corresponds to the answer given by my proposal.

As a final example consider a situation where all of our beliefs are called into question. Suppose you come to believe that all your belief-producing mechanisms are unreliable, due to some malfunction in your brain. Examples might be mad cow disease, LSD, or other drugs and illnesses. In such a situation all of your beliefs are called into question, since all of them proceed from your brain, which is affected by the disease or chemical. Intuitively the probability of our cognitive faculties being reliable in this situation is low. However, if in addition to this malfunction and our experience we were to conditionalize on the truth of our other beliefs, it would turn out that our cognitive faculties are probably reliable, which is the wrong result. To guard against this possibility I proposed excluding conditionalizing on the truth of any belief based on the source in question. Since the source of all of our beliefs is the malfunctioning brain, we should not conditionalize on any of our beliefs in this case. But we must conditionalize on all information that is not suspect, which would include a description of our experience.[7] This case is very similar to the situation involving naturalism in Plantinga's argument; in both cases we are considering whether the source of our beliefs is unreliable in all areas. And in both cases we cannot conditionalize on the truth of our beliefs without begging the question in an important sense.

In each of these examples we have a situation in which beliefs based on a certain source are called into question. And in each of these cases we see that we should not conditionalize on beliefs that are "tainted" by or based on that source. To do so would result in probabilities that are not useful for the arguments in which they are employed. Applying this to Plantinga's argument, we

[7] As mentioned earlier, we are not here assuming that the agent's beliefs about his experience are reliable. In applying Plantinga's argument the agent makes a pragmatic assumption about his experience, but these beliefs can be defeated.

should conditionalize on naturalism, evolutionary theory, and a description of our beliefs and experience. We should not conditionalize on the truth of our beliefs, since this is what is being questioned. We thus see that the naturalist is correct in objecting that we need to conditionalize on more than N&E in this argument; the additional information is a description of our experience. But we cannot conditionalize on information many have proposed, since that would be conditionalizing on the truth of beliefs proceeding from a source under question.

Is the Probability of R on Naturalism High?

This proposal of what to conditionalize on has immediate consequences for Plantinga's argument. Plantinga gave several examples that show how our beliefs could be adaptive but systematically wrong. From this Plantinga argued that the probability that our cognitive faculties are reliable given N&E was either low or inscrutable. Recall Plantinga's example of a person who believes everything is a witch and all his ways of referring to things involve definite descriptions entailing witchhood. Although Plantinga is correct that this is possible and shows that our beliefs could be adaptive and mostly false, having these beliefs is not consistent with our experience. Our experience is that our beliefs do not generally involve definite descriptions entailing witchhood. Thus we exclude such possibilities and they do not provide direct support for P(R/N&E&EXP) being low or inscrutable. Furthermore, we note that we have internal experience that we do not have any of the strange belief-desire combinations that Plantinga used in his original argument. In Plantinga's example of prehistoric Paul, Paul does not think the tiger is a pussycat, that he wants to pet it, and that the best way to pet it is to run a way from it. So in looking at the probability of R we should exclude worlds in which there are these unusual belief-desire combinations. We limit ourselves to worlds in which the beings' beliefs are the same as ours.

However, even though we exclude worlds in which we have adaptive but systematically wrong beliefs when determining P(R/N&E&EXP), we can use those worlds to develop an argument that P(R/N&E&EXP) is not high. The unusual beliefs Plantinga describes are not to be used in evaluating how likely R is, given naturalism and our experience, but they do affect that probability and provide us with a way of showing that it is not high. This unexpected result arises because even though the unusual beliefs Plantinga describes are inconsistent with our experience of our beliefs, they are true in worlds consistent with our external experience of the world. For example, in one of Plantinga's examples Paul has the external experience of seeing a tiger in front of him. This experience is consistent with his believing the animal is a big pussycat instead of a tiger, and it is also consistent with his believing the tiger signals the start of a footrace. Although these alternative beliefs Plantinga de-

scribes are inconsistent with Paul's internal experience of what he believes, they are consistent with his external experience, which does not include experience of his beliefs and desires.

We thus have a large number of sets of adaptive beliefs, including our normal beliefs and the unusual ones that Plantinga describes, that are consistent with the same external experience. If we were more formally minded, we might call this an equivalence class of beliefs for some external experience. The intuitive idea is that these belief sets are all consistent with the same external experience. But corresponding to each of these belief sets in that equivalence class will be a state of the world. Subtract the beliefs and desires of the agent from these states of the world, so that the states of the world include our external experience of the world, the external world itself, but no beliefs or desires of the agents in question. The resulting states of the world are consistent with all of our experience, both the internal experience of our beliefs and the external experience of the external world. In addition, these states of the world are inconsistent with each other; they describe different ways the world could be and still give rise to the same external experience. Thus for each of the unusual adaptive beliefs Plantinga describes there is a corresponding state of the world consistent with our total experience that would make the unusual beliefs true and our normal beliefs false. The unusual adaptive beliefs imply the existence of states of the world consistent with my experience, in which my beliefs are false but are adaptive and have survival value.

Furthermore, given naturalism, evolutionary theory, and my experience, there is no reason to think that any one of these states of the world is more likely to be actual than the others. There is no God who is working to make sure our beliefs are mostly true, and thus any of those states of the world are as likely as any other. The only information I have is naturalism, evolutionary theory, and my experience, but each of those states of the world is consistent with that information. Symmetry considerations preclude me from assigning the state of the world corresponding to our normal beliefs a higher probability than the others. But then it is not likely that the state corresponding to our beliefs is the actual one, and thus it is not likely that our beliefs are mostly true. Thus we see that the probability that our beliefs are reliable, given naturalism, evolutionary theory, and our experience, is not high.

To illustrate, consider some of the examples that Plantinga mentions. Let B_1 be the belief set that includes Paul's belief that the tiger is a large cuddly pussycat that he wants to pet and that the best way to pet it is to run away from it. Let B_2 be the belief set that includes Paul's belief that he is about to take part in a sixteen-hundred-meter race and the tiger is the starting signal. Like Paul's actual beliefs, both B_1 and B_2 are consistent with Paul's experience of the external world. Corresponding to B_1 and B_2 are states of the world that make the corresponding beliefs true and Paul's actual beliefs false. Delete Paul's beliefs from the states of the world corresponding to B_1 and B_2. These states of the world do not include any of Paul's beliefs, and so they are

consistent with Paul's internal experience as well. Consider the state of the world such that B_1 is true. This is a world consistent with naturalism, evolutionary theory, and Paul's experience. In this state of the world Paul has the same external experience he has in the actual world, but Paul's belief that there is a tiger in front of him is false. In the world corresponding to B_2, the tiger signals the start of a race that Paul has entered; thus belief set B_2 is true and Paul's beliefs about the tiger are false. These are states of the world consistent with Paul's total experience in which Paul's beliefs in question are false. So Plantinga's examples allow us to construct alternative states of the world that are consistent with Paul's experience but inconsistent with his beliefs.

Conditionalizing on our experience allows us to rule out that we have various other unusual belief-desire combinations, but we still have an argument that the probability of R is not high. We know what beliefs, desires, and experiences we have. Admittedly we believe our beliefs are the true ones, but given naturalism and that our beliefs were selected for by natural selection, the world might be very different than what we believe. One of the many other belief sets that lead to our actions might be true, and our beliefs arose via natural selection even though they aren't true. Since there are so many different worlds in which our beliefs are false but adaptive, there is no reason to think that our beliefs are mostly true. Given the naturalistic story about the source of our cognitive faculties, and given the experience and beliefs we have, it is quite unlikely that our beliefs are mostly true. Conditionalizing on admissible additional information does nothing to change this. The probability of our cognitive faculties being reliable, given naturalism, evolutionary theory, and our experience, is not high.

We have seen that Plantinga's argument against naturalism emerges unscathed by one of the major objections to it. When investigated in detail, conditionalizing on additional information does not weaken the preliminary argument, although there is an inherent weakness built into that argument. We then presented a proposal of what to conditionalize on in Plantinga's main argument, and we saw that this proposal gave the intuitively correct result in analogous cases. When applied to Plantinga's argument, we saw that this proposal was inconsistent with one way of construing Plantinga's argument that our cognitive faculties are unlikely to be reliable if naturalism is true. But we then gave an argument that this probability is in fact not high when the appropriate information is conditionalized on. In this chapter I have concentrated on the role that probability plays in Plantinga's argument; his argument raises fascinating issues in many different areas of philosophy. But like an onion, intense study only reveals more layers that need to be further investigated.

PART IV

THE NATURE OF EPISTEMIC DEFEAT

The Illusion of Defeat

W. J. Talbott

In chapter 12 of *Warrant and Proper Function*,[1] Alvin Plantinga argues that it is not rational to believe the conjunction of the theses of philosophical naturalism and evolutionary biology, where evolutionary biology is understood to include the claim that human cognitive faculties evolved by natural selection. I focus on what he refers to as his *main argument*.[2] The main argument crucially depends on the notion of an *undercutting defeater*, which Plantinga takes over from Pollock. Plantinga applies the notion in ways that go beyond Pollock's account, however. I argue that Plantinga has misunderstood the role of undercutting defeaters in reasoning. I believe that his misunderstanding leads him to laud as rational a kind of cognitive rigidity that is far from epistemically desirable.

Following Plantinga, I let *N* stand for the relevant tenets of naturalism; I let *E* stand for the relevant tenets of evolutionary biology; I let *R* stand for the claim that human cognitive faculties are reliable on the whole; and I let *C* stand for a complex proposition that describes what cognitive faculties human beings have.[3] The conclusion of the main argument is the extreme skeptical conclusion that those who believe N&E&C can have practically no rational beliefs—at least no rational beliefs of theoretical logic, mathematics, science, and metaphysics, and, thus, that N and E and C themselves are not rational.

This essay developed from a discussion of Alvin Plantinga's paper "Naturalism Defeated," which he presented at a University of Washington Philosophy Department colloquium in February 1996. I have greatly benefited from discussion with Plantinga at that time and from further contacts afterward.

[1] New York: Oxford University Press, 1993; hereafter *WPF*.

[2] *WPF*, 229–237.

[3] *WPF*, 220.

The crucial move in the main argument is the claim that those who believe N&E&C have a certain kind of defeater, which Pollock calls an *undercutting defeater* (to be even more specific, a *reliability defeater*),[4] for at least some of their empirical beliefs (including N&E&C itself).[5]

Plantinga carries through his main argument for two alternate assumptions: (1) On the assumption that it is rational to hold that P(R/N&E&C) << ½ and (2) on the assumption that it is rational to be agnostic about P(R/N&E&C). If Plantinga's argument fails in the first case, it must also fail in the second. In this chapter I argue that it fails in the first case.

Limited to case (1), the main argument has two versions, a *global* argument based on judgments of *global reliability or unreliability* (the reliability or unreliability of all of our cognitive faculties or belief-producing mechanisms), to the conclusion that all of our cognitive faculties are probably unreliable and all of our beliefs are irrational, and one or more *local* arguments, based on judgments of *local reliability or unreliability* (the reliability or unreliability of particular cognitive faculties or belief-producing mechanisms, not all of them), to the conclusion that only some of our cognitive faculties are probably unreliable and only some of our beliefs are irrational. Plantinga makes both types of argument. Because the local argument requires weaker assumptions, I will focus on it. If the local argument fails, then the global argument must fail.

Before setting out the main argument that it is irrational to believe E&N&C, it is useful to consider an example that Plantinga uses to explain the role of undercutting defeaters in the main argument.

The wish fulfillment example. Suppose a theist comes to believe that religious beliefs generally are almost always the result of wish fulfillment. Suppose also that "she thinks the probability that wish fulfillment is reliable is low, and the probability that a belief should be true, given that it is produced by wish fulfillment not far from ½. Then she has a straightforward defeater for any of her beliefs she takes to be produced by wish fulfillment."[6]

The first thing to note about the wish fulfillment example is that the argument involved is local. In this case, the evidence bears only on the reliability of a *single* cognitive faculty or belief-producing mechanism, not on the reliability of *all* of one's cognitive faculties or belief-producing mechanisms. Consider an agent AP who believes that God exists (G). From Plantinga's discussion of the example, I take it that we should assume that AP has evidence X, which is evidence that wish fulfillment (cognitive faculty WF) is probably un-

[4] John L. Pollock, "Reliability and Justified Belief," *Canadian Journal of Philosophy* 14 (1984): 103–114.

[5] In an oral presentation at the University of Washington on February 16, 1996 ("Naturalism Defeated"), Plantinga simplified the main argument by substituting P(R/N&E) for P(R/N&E&C). Though I think this substitution is an improvement to Plantinga's argument, it does not help with the objections that I raise shortly, so I can safely ignore it here.

[6] *WPF*, 230.

reliable and she also believes that her belief G is based on wish fulfillment. Let *R(WF)* stand for the statement that AP's faculty of wish fulfillment is reliable. Then it seems that Plantinga is claiming that that AP can reason as follows:

(Step 1) P(R(WF)/X) << ½.[7]
(Step 2) X
(Step 3) I believe G based on WF.

At this point, Plantinga claims that AP has a defeater for her belief G. Does that mean that it is irrational for AP to believe G? Not necessarily. Plantinga claims only that "if that defeater remains itself undefeated and if she has no other source of evidence, then the rational course would be to reject belief in God."[8] I consider Plantinga's two additional steps in reverse order:

(Step 4) I have no other basis for believing G.

Clearly, if AP had a basis for believing G from a cognitive faculty other than wish fulfillment, and the other cognitive faculty was one that she believed to be reliable, then her belief that G would not be undermined by steps 1–3. Step 4 closes off this possibility.

(Step 5) Steps 1–4 are not themselves subject to defeat.

Plantinga claims that AP's belief G is irrational if her defeater for G remains itself undefeated. To be clear about what is going on, I suggest that we rephrase Plantinga's point to say that steps 1–4 give AP a *prima facie* defeater for G. Whether it is rational for AP to believe G depends on whether this prima facie defeater becomes an *actual* defeater. The only thing that can save AP's belief G from actual defeat is if AP has other beliefs which defeat one or more of steps 1–4 (beliefs which are not themselves defeated by other be-

[7] Actually, what I have labeled step 1 does not exactly fit what Plantinga says about the wish fulfillment example. As mentioned in the text, what Plantinga actually says is that "the probability that a belief should be true, given that it is produced by wish fulfillment not far from ½" (230). This statement would be better rendered as follows:

Step 1' P(G/X and I believe G based on WF) ≈ ½.

However, if I had replaced step 1 with step 1', the argument in the wish fulfillment example would not exactly parallel Plantinga's main argument. To maintain the parallel, I have substituted step 1 for step 1'. This substitution should be harmless, because if the argument of the wish fulfillment example did not go through for step 1 as I state it in the text, then that example would undermine rather than support Plantinga's main argument.

[8] *WPF*, 231.

liefs).[9] But if, as we assume in the present case, nothing defeats steps 1–4, AP would reach the final step of the argument:

(Step 6) My belief G is irrational.

I agree with Plantinga that in the example as described, AP's belief G is irrational, and if she were able to carry out the first five steps of the preceding argument, she should realize that her belief G is irrational, but I do not believe that Plantinga provides a satisfactory account of the reasons that AP should recognize her belief G to be irrational. As stated above, steps 1 and 2 would leave it open for a belief B to be defeated by any evidence X, no matter how partial or gerrymandered, that the cognitive faculty responsible for the belief B is unreliable. I believe that step 1 needs to be augmented to make clear that X includes all of AP's beliefs relevant to the reliability of the cognitive faculty of wish fulfillment. I explain this suggestion below.

The discussion of the wish fulfillment example is well chosen to set the stage for Plantinga's final presentation of the main argument, because it enables him to present the main argument in its strongest form, not as a global argument but as one or more local arguments. As a concession to his opponent, he can allow that evolution might have selected for reliable faculties of perception and memory.[10] All that he requires is the more plausible claim that evolution would not have selected for reliable cognitive mechanisms producing the beliefs involved in theoretical logic, mathematics, science, or metaphysics.[11] I am sure that he would not object to my adding theoretical metaphysics to the list. The same arguments that apply to theoretical logic, mathematics, and science apply at least as strongly to theoretical metaphysics.[12] Since N is a part of theoretical metaphysics and E and C are both parts of theoretical science, Plantinga would then be in a position to produce a defeater for each conjunct of N&E&C. It is convenient to refer to the cognitive faculties that are involved only in the production of theoretical beliefs—including beliefs in theoretical logic, math, science, and metaphysics—as the *abstract faculties* (AF). Here then is my reconstruction of the main argument as applied to E (a parallel argument applies to N and C):

[9] Clearly, a full account of prima facie and actual defeat would require a careful recursive definition. My comments here are simply intended to point to the need for such a definition. It is not necessary for my present purposes to present the full recursive definition.

[10] *WPF*, 232.

[11] Ibid.

[12] There is an important difference between theoretical metaphysics, on the one hand, and theoretical logic, mathematics, and science on the other hand. Given the large amount of disagreement in theoretical metaphysics, a case can be made that human beings really are unreliable metaphysicians. Of course, this argument is independent of any information about our evolutionary heritage. Kant made a similar argument about the status of metaphysics long before Darwin. I set aside this issue here. For the purposes of the present argument, the reader should imagine that theoretical metaphysics is generally regarded to be as successful as theoretical logic, mathematics, and science really are regarded to be.

(Step 1) P(R(AF)/N&E&C) << ½ (that is, it is very improbable that human abstract faculties would be reliable, given that they have evolved by natural selection). I grant Plantinga this premise for the sake of the argument.
(Step 2) N&E&C (The argument is addressed to someone who accepts philosophical naturalism (N), evolutionary biology (E), and some agreed description of our cognitive faculties (C).)
(Step 3) I believe E based on AF.
(Step 4) I have no other basis for believing E (i.e., no other basis that does not make use of the abstract faculties).
(Step 5) Steps 1–4 are not themselves subject to defeat.
(Step 6) My belief E is irrational.

There is one obvious disanalogy between the wish fulfillment example and this argument. In the wish fulfillment example, the beliefs X which were the basis for doubts about the reliability of the faculty of wish fulfillment were not themselves produced by the faculty of wish fulfillment (but presumably by other faculties that the subject believed to be reliable). But in this case the beliefs N and E and C, which are the basis for believing the abstract faculties to be unreliable, are themselves the product of the abstract faculties. Thus, these beliefs seem to defeat themselves! It is not at all obvious what one should do when one has beliefs that defeat themselves. Plantinga cites Hume for a description of the kind of endless loop of apparent defeat followed by apparent rescue followed again by apparent defeat followed again by apparent rescue, etc.[13] Plantinga believes that there is no way to get out of the loop in a way that makes the beliefs N, E, and C rational.

Most significantly, Plantinga insists that it will do no good to try to bring in other reasons for thinking that the abstract faculties are reliable. Plantinga thinks that any such argument would have to be based on the frequency of true beliefs produced by the abstract faculties, but any such argument would obviously have to employ the abstract faculties themselves.[14] Plantinga argues that this move is no more philosophically respectable than the corresponding move by the theist (AP) to defend belief G by using wish fulfillment to gauge its own reliability. Plantinga claims that such an argument would be *pragmatically circular*.[15]

I think that Plantinga's conclusion in the wish fulfillment example is correct, but that he is mistaken about pragmatically circular reasons in general. He is correct that it cannot be rational to judge wish fulfillment reliable on the basis of wish fulfillment itself, *when there are other reasons (based on other cognitive faculties) for thinking it unreliable*. But when our cognitive faculties, individually or collectively, judge themselves to be reliable on the basis of the total evidence available, and there is no other basis for judging them to be un-

[13] *WPF*, 234–235.
[14] Ibid., 233.
[15] Ibid., 234.

reliable, I think it is a mistake to claim that that result is objectionably circular. Indeed, if one considers the alternatives (that they judge themselves to be unreliable or that they make no judgment about their reliability), judging themselves to be reliable provides the strongest kind of reason for trusting them. Of course, it is possible to imagine very unreliable faculties that would judge themselves reliable. That only shows that their judging themselves to be reliable does not guarantee that they are reliable. But it does not change the fact that, considering the alternatives, their judging themselves to be reliable is some reason to believe that they are reliable, and a better reason than either of the other two alternatives (viz., that they judge themselves unreliable or that they make no judgment about their reliability). I believe that Plantinga would disagree with me on this. I take up this disagreement shortly.

As I see it, Plantinga's argument is roughly that steps 1–6 of the argument in the wish fulfillment example have the same logical structure as steps 1–6 in the main argument, and that directly corresponding to the reasons that the theist should recognize that her belief G is irrational in the former case, the naturalist evolutionist should recognize that his belief E&N&C (and each individual conjunct) is irrational in the latter case. I think that there are important disanalogies between the two cases. In order to identify them, I construct a series of examples.

Example 1: the genetic hallucinogen blocker. There is a hallucinogen H so powerful that in 95 percent of subjects one dose of H causes severe hallucinations within one hour of taking it. Let *HCH* (H causes hallucinations) stand for the previous statement. BT (Bill Talbott) believes HCH. One hour ago BT swallowed a pill that he believes to have been one dose of H. BT also believes *ISH* (I swallowed H). Given only HCH and ISH, BT believes it is very improbable that his perceptual faculties are reliable (i.e., very improbable that *R(PF)*). BT currently has the perceptual belief that he is typing on his computer (*TC*). Suppose that BT recognizes that his belief TC is the product of his perceptual faculties (PF). BT seems to be in a position to go through the six steps that lead to the conclusion that his belief TC is irrational:

(Step 1) P(R(PF)/HCH & ISH) = .05 << ½.
(Step 2) HCH & ISH
(Step 3) I believe TC based on PF.
(Step 4) I have no other basis for believing TC.
(Step 5) Steps 1–4 are not themselves subject to defeat.
(Step 6) My belief TC is irrational.

However, even though BT can legitimately take steps 1–5, I do not believe that it is always legitimate for BT to take the final step. Suppose, for example, in this case, BT knew, in addition, that 5 percent of the population had a gene which produced a protein that blocked the effects of drug H. I abbreviate that proposition as *5GBH* (5 percent genetic block to H). Also, suppose that BT

knew that he himself had that genetic block (*IHGBH*—I have the genetic block to H). Suppose also that he has plenty of evidence, Y, that his perceptual faculties are reliable when he is not under the influence of hallucinogenic drugs. Then, in light of BT's total evidence, he would also be able to take the following step:

(Step 1-1) P(R(PF)/HCH & ISH & ~GBH & IHGBH & Y) >> ½

What has long been noted in discussions of probability is that step 1 and step 1-1 do not contradict each other. They can both be correct, because they give the probability of R(PF) based on two different reference classes (or properties). However, if both of them can be true, it is important somehow to restrict their use in reasoning, so that a subject does not come to the conclusion that it is both probable (based on step 1-1) and not probable (based on step 1) that her perceptual faculties are reliable. There is no generally accepted solution to what is referred to as the problem of *direct inference* for probabilistic inference.[16] But there is general agreement that a narrower reference class based on more evidence (as in step 1-1) takes priority over a broader reference class based on less evidence (as in step 1).

Earlier, I suggested that step 1 be qualified to require that it be based on the subject's total evidence concerning the reliability of the relevant cognitive faculty. This example illustrates why I made that addition. I believe that it is not irrational for BT to believe TC, given all the relevant evidence that he has concerning the reliability of his perceptual faculties (including the fact that he has a genetic blocker to the hallucinogenic effects of H). This is one important disanalogy between Plantinga's wish fulfillment example and his main argument. In his wish fulfillment example, it is clear that we are to suppose that AP's *total evidence* supports the conclusion that wish fulfillment is an unreliable belief-forming mechanism. But in the main argument, Plantinga depends on being able to consider only N&E&C in the determination of the reliability of the abstract faculties, not the total evidence relevant to their reliability.

I believe that Plantinga might well agree with me about what it is rational for BT to believe in example 1, and thus that there is some sort of Total Relevant Evidence Requirement on step 1 of his main argument. Whatever Total Relevant Evidence Requirement Plantinga might try to formulate, I believe that his strictures on pragmatic circularity will prevent it from being adequate. To understand why, consider the following series of variations on example 1.

Example 2: first example of the tardy revelation. This example is like example

[16] The best account of direct inference I know of is that of John L. Pollock, *Nomic Probability and the Foundations of Induction* (Oxford: Oxford University Press; 1990), chap. 4. I believe that even Pollock's account has problems, though it would take me too far afield to discuss them here.

1, in that BT has the gene that blocks the effects of drug H, but differs from example 1 in that BT himself has no information on whether or not he has this gene. It has now been just over an hour since BT took the drug H. BT reasons that if he does not have the drug-blocking gene, he is now under the influence of the drug. He believes that it is probable that his perceptual faculties are unreliable, but he has to admit that he has not noticed much change in his perceptions. He wonders if the effects of the drug only manifest themselves gradually. Just then his doctor calls him on the phone and tells him that a genetic test has been performed and the results show that he has the H-blocking gene.

In this example, I think we can agree that it was rational for BT to expect his perceptions to be unreliable until he got the phone call from his doctor. My position is that after he got what he took to be a phone call from his doctor reporting that he had the H-blocking gene, he was in very much the same position as BT in the first example above. His total evidence would not rationally support doubts about the reliability of his perceptual faculties. When he puts down the phone and goes back to work on his computer, he acquires the perceptual belief that he is typing at his computer (TC). Because, given his total relevant evidence, it is rational for him to believe that his perceptual faculties are reliable, his belief TC is rational.

It seems to me that Plantinga would have to disagree. On Plantinga's account, once it was rational for BT to expect his perceptions to be unreliable, it would be irrational for BT to rely on perceptual evidence (in this case his seeming to have a phone conversation with his doctor in which the doctor told him that he had the H-blocking gene) to justify the conclusion that his perceptual faculties were not reliable. In contrast, I see nothing wrong with BT's reasoning in this case. Although BT would have to consider the possibility that the apparent phone call from his doctor was really only a hallucination, it seems to me that, in light of the apparent absence of other hallucinatory effects on his experience and in view of the fact that it could only seem an incredible coincidence for the drug H to cause the hallucination of his doctor calling to tell him that he had the H-blocking gene, it would be rational for BT to believe that the doctor's call was not a hallucination. Thus, I believe that in this case there is nothing objectionably "circular" about BT's use of his perceptual faculties as sources of evidence for the beliefs which in turn support his belief that he is not under the influence of the drug and that his perceptual faculties are reliable. I think that the main disagreement between Plantinga and me is over which sorts of use of our cognitive capacities to determine their reliability are objectionably "circular" from an epistemic point of view. To focus the disagreement even more precisely, I continue the examples.

Example 3: the antidote. Suppose the facts in example 2 up to the point that BT ingested the drug H, except that in this example, BT does *not* have the H-blocking gene. In this case, one hour after ingesting the drug, BT begins

to have very weird perceptions—pink elephants dancing, etc. The drug is so effective that BT believes that he is really seeing pink elephants dance. Suppose that before he fell under the influence of the drug, BT knew that it exerted its effects for twenty-four hours. However, he did not know that there was an antidote. Six hours after BT ingests the drug, a friend of BT's happens by. (They had not planned to get together, and earlier BT would have regarded a visit from any of his friends as highly unlikely.) Alarmed at BT's description of his perceptual world, the friend drives him to a hospital emergency room where the antidote to H is administered. It is quickly effective and BT's perceptions go back to what seems to him to be normal (and are in fact normal). He is told that he took drug H eight hours ago, that an antidote was administered, and that the antidote was effective. Then his friend drives him home and he goes back to work at his computer, thus acquiring the perceptual belief that he is typing at his computer (TC).

Even though earlier BT had rationally believed that the drug would make his perceptual faculties unreliable for twenty-four hours and even though, when he took the drug, he had no idea that a friend would stop by and take him to the emergency room, or even that there was an antidote to H, it seems to me that it could well be rational for BT to believe that he had received an antidote to H that restored the reliability of his perceptual faculties, and thus rational for BT to believe TC—all based, at least in part, on evidence from his perceptual faculties. Again, it seems to me that Plantinga would have to claim that BT's reasoning is objectionably "circular." So we would disagree about this case also.

In the remainder of this chapter, I address potential disanalogies between my examples and Plantinga's main argument. First, Plantinga's argument seems to have a temporal structure lacking from my examples. Plantinga seems to be arguing that there was a time (before humans even appeared on the earth) at which it was very improbable that creatures with reliable abstract faculties (or cognitive faculties generally) would have evolved. To capture this part of Plantinga's argument, I need to employ David Lewis's distinction between subjective probability and objective chance.[17] Second, my examples involve only one person; but Plantinga's argument applies to all human beings, because according to Plantinga, it is probable that evolution would have made all of us unreliable (at least in our theoretical judgments). The following example corrects for these two disanalogies:

Example 4: HS invades the food chain. Suppose chemical compound HS has entered the food chain. All human fetuses throughout the world have been exposed. HS would have made all children in the world incurable hallucinatory schizophrenics, had not there been a random mutation in a benign bacteria that enabled it to digest compound HS. Before the mutation, its objec-

[17] David Lewis, "A Subjectivist's Guide to Objective Chance," in *Studies in Inductive Logic and Probability*, vol. 2, ed. R. C. Jeffrey (Berkeley: University of California Press, 1980), 263–293.

tive chance of occurring was very low. After the random mutation, the benign bacteria was cultured and a small dose was given to every child in the world. No child actually developed hallucinatory schizophrenia from exposure to compound HS, even though the objective chance was high, when they absorbed compound HS, that they would develop it. On his twenty-first birthday, BT is informed of the lucky mutation that saved him from becoming an incurable hallucinatory schizophrenic as a child. As he sits down at his computer and begins typing an email message, does he have an undercutting defeater for his belief TC? My understanding of Plantinga's main argument is that BT would, at age twenty-one, have a defeater for his perceptual beliefs, including TC, because when he was *in utero* there was a high objective chance that he and all other human beings would become incurable hallucinatory schizophrenics. I do not see how it could be rational for BT to reason in this way. I believe that BT has lots of evidence of the reliability of his perceptual faculties that would block Plantinga's argument.

Another disanalogy between my examples and Plantinga's argument is that my examples concern perceptual beliefs, while Plantinga's argument concerns theoretical beliefs. This should work against my examples, because evidence of reliability will almost always be, at least in part, perceptual. Thus, beliefs about the reliability of one's perceptual faculties ought to be the ones that would be most liable to problems of 'pragmatic circularity'. Indeed, my examples all illustrate 'pragmatic circularity', but it seems to me that the "circularity" is not problematic from an epistemic point of view. However, it would be possible to change any of the four previous examples, so that the drug H affected the reliability of theoretical beliefs (whether of logic, mathematics, science, or metaphysics). I refer to these as the *modified examples*. If BT were a competent logician, mathematician, scientist, or metaphysician, I would not expect his logical, mathematical, scientific, or metaphysical beliefs to be defeated in any of the four modified examples.

I will close this chapter by saying why I think that Plantinga's account of defeat is defective. On Plantinga's account, rationality exhibits a kind of rigid dependency on the past. Judgments that one's cognitive faculties were probably unreliable would be irreversible, if the only evidence one might obtain of their reliability required the use of those very faculties. But rationality requires more flexibility than this. A rational person is always willing to entertain new hypotheses, even antecedently improbable ones, in the light of new evidence. Even the belief that one's perceptual faculties are unreliable can be rationally revised in the light of appropriate perceptual evidence. Because Plantinga's account would rule out such flexibility, it turns out to be an account of a kind of cognitive inflexibility rather than of rationality. On Plantinga's account, once one had judged that one's cognitive faculties were probably unreliable, it could never be rational to use those cognitive faculties to reconsider the earlier judgment.

It does seem to me that it is *possible* for someone to find herself in the kind

of self-defeating position that Plantinga thinks those of us who believe N&E&C are in, but it is much more difficult than he suggests to get into such a position. Consider the following variation on example 2 above:

Example 5: second example of the tardy revelation. In example 2, BT had the experience of the phone ringing and his doctor telling him that he had the H-blocking gene. Suppose that at the time he took the drug H, BT had realized that H typically produced just that sort of hallucination. Then the experience of his doctor calling would *confirm* that he was under the influence of the drug and hallucinating. Even if his doctor had in fact called him, it would not be rational for BT to believe that the doctor had called him, because given what he knew about the effects of H and the probability of his being susceptible to the drug, it was much more probable that he was under the influence of the drug and hallucinating than that he had really been called by his doctor.

Of course, if there were a drug that caused one to hallucinate one's doctor calling to say that one was not susceptible to the effects of the drug, presumably doctors would find some other way to inform their patients of their immunity than by calling them on the telephone! In any case, this example would help Plantinga only if he were in a position to argue that, given N&E&C, it is probable *both* that we would have experiences of the kind that we generally have *and* that at least our abstract faculties—if not all of our cognitive faculties—would be unreliable. I cannot imagine how facts about evolution could have made it probable that we would have the sorts of experiences that we do have, unless our perceptual *and* theoretical beliefs were anchored by a world that they describe to at least a good approximation.

For example, in theoretical logic, mathematics, science, and even metaphysics, there are numerous examples of situations in which experts' preconceptions about what must be the case have been disappointed, and the experts have felt compelled to alter their beliefs to accommodate the new discoveries.[18] I believe that this is evidence that, like our perceptual beliefs, our theoretical beliefs run up against objective constraints. In combination with the other evidence that we have, I believe that this evidence of running up against objective constraints constitutes powerful evidence of the long-term reliability of what I have called our abstract faculties. It is not merely that logicians, mathematicians, scientists, and metaphysicians agree (to the extent that they

[18] In logic, Godel's incompleteness theorem (showing the incompleteness of any finite axiomatization of arithmetic) had this effect. Before Godel's result was announced, most logicians expected a proof of the *completeness*, not the *incompleteness*, of the arithmetical axioms. In mathematics, Cantor's work on transfinite arithmetic had this effect. Before his work, mathematicians were reluctant even to employ the concept of infinity. There are many examples from theoretical physics—for example, Einstein's special theory of relativity showed that a generally shared assumption among physicists about the nonrelativity of simultaneity was mistaken. Examples in metaphysics are more controversial, but I believe that there are such examples in metaphysics also—for instance, Aristotle's example of the ship of Theseus showed that a general assumption about identity through time was mistaken.

do) that gives us evidence of the long-term reliability of our abstract faculties, but rather that their agreement itself seems to be a response to objective constraints rather than simply a group reinforcement of prior expectations. It seems to me that this gives us good reason for judging our abstract faculties to be reliable in the long term. Of course, this very argument for the long-term reliability of our abstract faculties is a highly theoretical argument. But I see no objectionable circularity in relying (at least in part) on our abstract faculties to determine their long-term reliability. As discussed above, it is inevitable that some determinations of the reliability of our cognitive faculties will be "circular" in this sense. When one considers the alternatives, this sort of "circularity" cannot be considered a vice.

So I conclude that Plantinga's main argument fails. It fails because the conception of rationality embedded in Plantinga's argument is not rationality at all, but rather a quite irrational cognitive rigidity that is far from desirable from an epistemic point of view.

Conditional Probability and Defeat

Trenton Merricks

Here is an outline of Alvin Plantinga's "Evolutionary Argument against Naturalism."[1] The probability of one's own cognitive faculties' being reliable (R specified to oneself)[2]—given both metaphysical naturalism (N) and evolution (E)—is low or inscrutable.[3] That P(R specified to oneself/N&E) is low or inscrutable gives the devotee of N&E a defeater for R specified to herself.[4] That defeater is undefeated. And that implies that the evolutionary naturalist has a defeater for all her other beliefs, including N.[5]

Thanks to Michael Bergmann, Eugene Mills, Alvin Plantinga, Michael C. Rea, and Donald P. Smith.

[1] The original version of that argument appears in Plantinga's *Warrant and Proper Function* (New York: Oxford University Press, 1993), 216–237 (hereafter *WPF*), and is further developed in his "Naturalism Defeated" (unpublished).

[2] Plantinga is not always explicit about the fact that he is arguing that the evolutionary naturalist has a defeater for R *specified to herself*. But Plantinga's conclusion that the evolutionary naturalist's *own beliefs*—including her belief in N—are subject to an undefeated defeater follows only if R specified to herself is defeated.

[3] Plantinga says we could think of the relevant probability as either *epistemic* or *objective*, although he favors the latter (*WPF*, 220 n. 7). Nothing in this chapter depends on taking the probability here one way rather than the other.

[4] I leave open whether the alleged defeater for R is N&E alone or N&E conjoined with the belief that P(R/N&E) is low or inscrutable. For stylistic reasons I will, however, occasionally say things like "Plantinga thinks N&E defeats R."

[5] Fred Dretske ("Epistemic Consequences," *Journal of Philosophy* 67 [1970]: 1007–1023) and Robert Nozick (*Philosophical Investigations* [Oxford: Clarendon Press, 1981]) argue that one's knowledge of particular bits of fact—such as that one is sitting in a chair—is consistent with one's not knowing whether one is a brain in a vat. But Plantinga argues that if one is not reasonable in believing R with respect to oneself, one has an undefeated defeater for all one's other beliefs. Similarly, if one is not reasonable in believing that one is not a brain in a vat, Plantinga's arguments would seem to show that—pace Dretske and Nozick—one has an undefeated defeater for everything else one believes; and so thoroughgoing skepticism follows.

In this chapter, I will focus on only one claim, a claim at the very heart of the Evolutionary Argument:

> The Crucial Claim: If P(R specified to oneself/N&E) is low or inscrutable, then one who believes N&E has a defeater for R specified to herself or himself.

If the Crucial Claim is not true, the Evolutionary Argument is unsound.

I

One might think that the Crucial Claim is obviously and sort of trivially true. For one might mistakenly endorse:

> The Obviously False Claim: (p)(q) If P(p/q) is low or inscrutable, then one who believes q has a defeater for p.

The Obviously False Claim's *virtue* is that accepting it would allow us to conclude right away that the Crucial Claim is true. But its *vice* is its truth value. Consider:

> p = I was thinking about Plantinga's evolutionary argument yesterday afternoon
> q = the sun is larger than the earth

I suppose P(p/q) is low. Or perhaps it is better to say it is inscrutable. Yet I am still warranted in believing that p. For there is no defeater for p anywhere in the neighborhood of q. Note that if q *were* a defeater for p, it would presumably be an *undefeated* defeater (and so p would be irrational for me, which it isn't). After all, I don't have any defeater for the claim that P(p/q) is low/inscrutable; nor do I have a defeater for the claim that q. I happily embrace both claims, and recognize that there is no need to defeat either in order to preserve the rationality of p. For they are irrelevant to the rationality of my believing that p.

The Obviously False Claim is not more or less on target, right in its general idea, only a little chisholming away from perfection, vulnerable only to cleverly contrived counterexamples. The Obviously False Claim is utterly misguided and confused. Note that my counterexample is not particularly clever. It involves no funny trick, no funny relation between purported defeatee and wannabe defeater that blocks the normally justified inference from low or inscrutable conditional probability to defeat. Quite the contrary. Counterexamples of this sort can be manufactured at will.

Imagine I'm dealt the following poker hand: seven of clubs, ace of hearts, six of diamonds, queen of spades, and six of spades. I see these cards in my hand; I'm warranted in believing I have been dealt exactly those five cards. Is

this belief defeated because I *also* know that being dealt exactly that hand is wildly improbable on many other things—indeed on virtually every other thing—I know? Certainly not. Suppose again I am warranted in my memory that I once presented—with willful and shameless disregard for the word limit—a paper at the Central APA that was exactly 3,004 words long. Suppose someone tells me that the probability of my presenting a paper of exactly that length at the APA is low or inscrutable on the fact that I am under seven feet tall. Does my memory belief thereby have a defeater? Not even close.

I emphasize the falsity and complete confusion of the Obviously False Claim not as an objection to Plantinga, for he explicitly rejects the Obviously False Claim.[6] Rather, I emphasize it because I want us to see that, in general, inferences from low or inscrutable conditional probability to defeat are unjustified. Once we see that, we should also see that anyone who endorses this sort of inference in a particular case owes us an explanation as to why, in that case, the inference is justified. At least, such an explanation is owed unless that case is one in which it is intuitively obvious that the inference is justified. But I do not think the Crucial Claim is intuitively obvious. So the burden of proof is on one who would endorse it.

II

I think Plantinga would agree that he should not simply assume the Crucial Claim. And he does, in fact, *argue* for it. His argument involves *analogies*. He suggests cases where low or inscrutable conditional probability seems to lead to a defeater, and then suggests that the evolutionary naturalist's position is, regarding R specified to herself, relevantly analogous to those cases. Thus, he argues, those cases support the Crucial Claim.

Note that we won't get support for the Crucial Claim from just *any* case of a low or inscrutable P(p/q)'s implying that q is a defeater for p. For the mere fact—if it is a fact—that there are cases in which a low or inscrutable P(p/q) implies that q is a defeater for p shows only that *some* instances of the Obviously False Claim are true. But this alone no more supports the Crucial Claim than it does any and every other instance of the Obviously False Claim. That is, it does not support the Crucial Claim at all.

So if Plantinga is to muster support for the Crucial Claim by way of analogies, those analogies must involve more than the fact that the defeatee is improbable or inscrutable on something else one believes. The defeatee in the analogy must, in some further and deeper way, be relevantly epistemically analogous to the evolutionary naturalist's belief in R specified to herself. For this reason, I will focus my discussion on Plantinga's analogies that seem most

[6] "Naturalism Defeated," 38–40.

likely to be of sufficient depth, those that involve an alleged case of *R's being defeated*.[7]

So—to consider one of the central analogies Plantinga invokes—suppose that all we know about the members of a hypothetical population is that their cognitive faculties came about as a result of naturalistic evolution. Suppose also that we accept that the probability of those faculties' being reliable, given that they came about by purely naturalistic processes, is low or inscrutable. Plantinga then asks:

> What would then be the appropriate attitude toward R (specified to that hypothetical population)? Someone who accepts N&E and also believes that the proper attitude toward P(R/(N&E))[8] is one of agnosticism clearly enough has good reason for being agnostic about R as well. She has no other source of information about R (for that population); but the source of information she does have gives her no reason to believe R and no reason to disbelieve it. The proposition in question is the sort for which one needs evidence if one is to believe it

[7] Because an analogy that supports the Crucial Claim requires more than defeat via conditional improbability or inscrutability, we can ignore any analogies that don't have the requisite "more," such as, for example, the case involving widgets (*WPF*, 230–231). What, exactly, *is* required for a defeatee to be relevantly analogous to the evolutionary naturalist's belief in R specified to herself? It's hard to say. Maybe the defeatee's involving a *belief-forming faculty* is enough. For that reason, I'll address the following analogy in this note:

> Compare the case of a believer in God, who, perhaps through an injudicious reading of Freud, comes to think that religious belief generally and theistic belief in particular is almost always produced by wish fulfillment. Such beliefs, she now thinks, are not produced by cognitive faculties . . . aimed at truth; instead, they are produced by wish fulfillment, which . . . does not have the function of producing true beliefs. Suppose she considers the objective probability that wish fulfillment, as a belief-producing mechanism, is reliable. [Suppose she concludes that probability is low or inscrutable.] But then . . . she has a defeater for any belief she takes to be produced by the mechanism in question. . . . But now suppose we return to the person convinced of N&E who is agnostic about P(R/[N&E]); something similar goes for him. He is in the same position with respect to any belief B of his, as is that believer in God. (Plantinga, *WPF*, 229–231)

Note that in the wish fulfillment case, one *first* ceases to believe that a certain faculty—in this case, wish fulfillment—is "aimed at truth." Once one has reached that conclusion, Plantinga argues, one then has a defeater for any belief produced by that faculty.

This case would be analogous to the evolutionary naturalist who *ceased* to believe that her cognitive faculties were aimed at truth. But Plantinga is not entitled to assume that the evolutionary naturalist has ceased to believe this. (Presumably anyone who accepts R specified to himself will accept that his cognitive faculties are, unlike wish fulfillment, in some important sense aimed at truth.) And note that the evolutionary naturalist's accepting that P(her faculties' being aimed at truth/N&E) is low or inscrutable is a far cry from her ceasing to believe that they are, in fact, aimed at truth. It is also a far cry from her having a defeater for their being aimed at truth. At least, one would have to *argue* that such a defeater follows; such an argument is likely to parallel the defense of the Crucial Claim itself.

[8] In the passage quoted, Plantinga is actually concerned with P(R/N&E&C), where C is a proposition setting out some of the main features of our cognitive system. But in "Naturalism Defeated" he drops C because he now thinks "the additional complexity unnecessary" (5). To avoid needless complexity, I too will drop C and will even replace 'N&E&C' with 'N&E' in quoted passages.

> reasonably; since there is no evidence the reasonable course is to withhold belief.[9]

Plantinga thinks that we should be agnostic about R specified to the hypothetical population. But this is not because R, specified to that population, is defeated. Perhaps surprisingly, his discussion of the hypothetical population contains no mention of defeaters. Rather, Plantinga recommends agnosticism because, first, there is no evidence for R specified to that population; and, secondly, we could be reasonable in believing R for that population only if we did so on the basis of evidence.

And as Plantinga shifts from the hypothetical population to the question of whether the evolutionary naturalist is reasonable in believing R specified to herself, his focus is still lack of evidence, not defeaters:

> But now suppose we again apply the same sort of reasoning to ourselves and our condition. Suppose we think N&E is true: we ourselves have evolved according to the mechanisms suggested by contemporary evolutionary theory, unguided and unorchestrated by God or anyone else. Suppose we think, furthermore, that there is no way to determine P(R/(N&E)) (specified to us). What would be the right attitude to take to R? Well, if we have no further information, then wouldn't the right attitude here, just as with respect to that hypothetical population, be agnosticism, withholding belief?[10]

Let's grant Plantinga what seems correct. Evidence is required for us to be reasonable in believing R with respect to that hypothetical population, evidence that we lack. And so, lacking that evidence, we are unreasonable in believing R specified to that population. But this would imply that the evolutionary naturalist is likewise unreasonable in accepting R specified to ourselves—or, more to the point, that the evolutionary naturalist is likewise unreasonable in accepting R specified to herself—only if belief in R specified to oneself *required evidence* to be reasonable.[11] But, by Plantinga's own lights, belief in R specified to oneself does *not* require evidence to be reasonable, for it is properly basic.[12] Thus, again by Plantinga's own lights, the case of the hypothetical population in no way impugns the evolutionary naturalist's belief in R specified to herself.

[9] *WPF*, 229.

[10] Ibid.

[11] 'Evidence' here means *propositional* evidence. Plantinga's argument involving the hypothetical population would not be strengthened by taking *non*propositional evidence into account. For the evolutionary naturalist—if R specified to herself is properly basic—presumably has nonpropositional evidence for R specified to herself, even though she lacks it for R specified to the hypothetical population.

[12] Cf. "Naturalism Defeated," 50–53, and *WPF*, 233–234.

III

There is another, more telling, objection to Plantinga's argument involving the case of the hypothetical population. For even setting aside the claim that R specified to oneself is properly basic, we should still agree that, prior to believing N&E, one who is now a devotee of N&E was *initially warranted* in believing R specified to herself.[13] That is, we should all grant that she was warranted—either in a properly basic manner or in some other way—in believing R specified to herself *before* believing those things that, according to Plantinga, generate a defeater.[14]

As a result, if N&E imply a defeater, N&E must dislodge the evolutionary naturalist from this initially favorable epistemic position regarding R specified to herself. There is, however, no initially favorable epistemic position regarding R specified to the hypothetical population. This crucial difference—and here is the "more telling" objection—undermines any relevant epistemic analogy between the hypothetical population and the evolutionary naturalist, and so undermines the hypothetical population's usefulness in defending the Crucial Claim.

Let's see if we can come up with a genuine epistemic analogy, inspired by the case of the hypothetical population. And we should try to construct our case so that it—unlike the original case of the hypothetical population—involves R's being *defeated*, thus increasing the likelihood that our case will be of use in justifying the Crucial Claim, itself a claim about defeat.

Consider the *revised* case of the hypothetical population. In the revised case, as in the original, one comes to believe that the members of a hypothetical population were produced by naturalistic evolutionary processes, and, moreover, that the probability of their faculties' being reliable, given that they were so produced, is low or inscrutable. But unlike the original case, in the revised case one is *initially warranted*—before the question of the population's origin ever occurred to one—in believing that the members of that hypothetical population have reliable cognitive faculties. Indeed, let's stipulate that, in the revised case, one is initially as warranted and as confident in believing R specified to the members of that population as one is in believing R specified to oneself.

[13] Certainly Plantinga should grant this. At least, it would be unfortunate if Plantinga's argument asked us to *start* with the claim that the evolutionary naturalist is not warranted in accepting R specified to herself, even before she comes to believe any alleged defeaters. And if we are entitled to assume that the devotee of N&E was never warranted in accepting R specified to herself, why bother *arguing* that R specified to herself is defeated?

[14] A sufficient condition for a belief B's being "initially" warranted "prior" to being defeated by D would be if B was actually warranted at some time temporally prior to a time at which, first, one comes to believe D and, secondly, D is a defeater for B. This the simplest case and I will consider cases of only this sort in the text. But there are surely atemporal notions of "initially" and "prior" lurking nearby, and these atemporal notions are probably the more fundamental and important ones (cf. "Naturalism Defeated," 53–56).

One's epistemic situation regarding R specified to the hypothetical population, in this revised case, does seem analogous to the evolutionary naturalist's epistemic situation regarding R specified to herself. But this is not to say that the revised case gives us a reason to endorse the Crucial Claim. For it is not clear—given our revisions—that one has a defeater for R specified to the hypothetical population.

Consider the fact that one is as *initially warranted* in believing R specified to the population as one is in believing R specified to oneself. Perhaps we should imagine that one has interacted with the members of that population and lived among them. (This seems a standard way to become warranted in believing R specified to some group of persons.) In such a case, it is not at all clear that R specified to the members of that population really *is* subject to a defeater when one comes to believe that they evolved in such a way that the probability of their being reliable, on their so evolving, is low or inscrutable.

To better see the point here, note that, for all I have said so far, the "hypothetical population" could be human beings other than oneself. So suppose one's initial noetic structure includes believing, with warrant and conviction, that other humans with whom one has interacted have reliable cognitive faculties. Suppose that one also comes to believe that these other humans have evolved by way of naturalistic processes, and moreover that P(R specified to others/N&E) is low or inscrutable. It's not *obvious* that one now has a defeater for others' reliability. Nor would anything relevant to defeat of R seem to change if we discovered that these "others" were actually alien impostors masquerading as humans. (It could happen.)

And, realistically, it's hard to imagine anyone who would accept that one has a defeater for R specified to others in the sort of case just noted who does not *already* accept that the evolutionary naturalist has a defeater for R specified to herself. For anyone who thinks that the revised hypothetical population provides one with a case of defeat will have to accept:

> The Analogous Claim$_1$: If P(R specified to others/N&E) is low or inscrutable, then one who believes N&E has a defeater for R specified to others.

But it seems that anyone who does not already accept the Crucial Claim will not accept the Analogous Claim$_1$, whether the "others" are one's kith and kin or a nonhuman race one has come to believe, with warrant, has reliable cognitive faculties. Any argument for the Crucial Claim that requires the Analogous Claim$_1$ is bound to be question-begging, or so nearly question-begging as to make no nevermind.

As originally presented in *Warrant and Proper Function*, the case of the hypothetical population is not epistemically analogous to the situation in which the evolutionary naturalist finds herself. Once the case is repaired so as to be analogous, it fails to be an *obvious* case of R's being defeated. (Remember, the Obviously False Claim is false.) Moreover, and perhaps more importantly, it

fails to be a case that will persuade one to accept the Crucial Claim who does not already accept it.

IV

Plantinga considers other analogies in defense of the Crucial Claim:

> Suppose I believe that I have been created by an evil Cartesian demon who takes delight in fashioning creatures who have mainly false beliefs (but think of themselves as paradigms of cognitive excellence): then I have a defeater for my natural belief that my faculties are reliable. Turn instead to the contemporary version of this scenario, and suppose I come to believe that I have been captured by Alpha-Centaurian superscientists who have made me the subject of a cognitive thought experiment in which the subject is given mostly false beliefs: then, again, I have a defeater for R.[15]

For the points I'll raise in this chapter, there is no difference between the Cartesian demon and the Alpha-Centaurian superscientists. I'll focus on the demon; what I say about it applies, mutatis mutandis, to the Alpha-Centaurians.

Suppose that I came to believe—somehow—that I was the product of a malicious Cartesian demon, a demon all of whose creations have mainly false beliefs and so, presumably, unreliable cognitive faculties. In such a case, I think Plantinga is right that I would be in some sort of epistemic trouble. Suppose further that in such a case I have a defeater for R with respect to myself. It is easy to see how the defeat is supposed to go here. I've come to believe that the demon's creations do not have reliable cognitive faculties. I also believe I am one of that demon's creations. And so I come to believe that I do not have reliable cognitive faculties. In the most straightforward way possible, I come to believe that R specified to me is false.

Plantinga cannot insist that the devotee of N&E reason to a defeater for R, specified to herself, in this straightforward way. For Plantinga does not suggest that his arguments show that the evolutionary naturalist must believe the very strong claim that the products of N&E are—not just probably but—*actually* unreliable. It is this very strong claim that is analogous to the claim that one is produced by an evil demon whose products actually have unreliable cognitive faculties. So, even granting Plantinga what he says about P(R/N&E)'s being low or inscrutable, the hypothesis that N&E is true is *not* analogous to the hypothesis that we are created by a demon who has given us unreliable faculties.[16] Here, as with the original case of the hypothetical population, the advertised epistemic analogy fails to hold.

[15] "Naturalism Defeated," 11.

[16] Plantinga also suggests one need not *believe* the evil demon hypothesis in order to have a defeater for R; rather, says Plantinga, R is defeated just so long as "*for all I know or believe* [the evil

If Plantinga wants to trade on an analogy between N&E and an evil demon story, he must replace the *standard* evil demon hypothesis with another. Let the *revised* evil demon hypothesis be that replacement, tailor-made for Plantinga's purposes. According to the revised demon hypothesis, we are all the creations of a revised demon. It is not part of that hypothesis that all of the revised demon's products are actually unreliable, but we do know that P(R specified to ourselves/revised demon hypothesis) is low or inscrutable. And—to preserve the central elements of the analogy with the evolutionary naturalist's belief in R specified to herself—according to the revised demon hypothesis, one is initially warranted in believing R specified to oneself.

If the revised evil demon hypothesis is to do its work for Plantinga, a believer of that hypothesis would have a defeater for R specified to himself. But while the *standard* evil demon hypothesis clearly and obviously places one in epistemic jeopardy, it is not so obvious that the *revised* evil demon hypothesis does. For it is not obvious that the following is true:

> The Analogous Claim$_2$: If P(R specified to oneself/revised demon hypothesis) is low or inscrutable, then one who believes the revised demon hypothesis has a defeater for R specified to herself or himself.

To get a better feel for what—besides the falsity of the Obviously False Claim—might make one suspicious of the Analogous Claim$_2$, imagine one is initially warranted in believing that *someone else* has reliable cognitive faculties, but then comes to believe the revised demon hypothesis as applied to that person. For example, I am now warranted in believing R specified to my brother. Suppose I come to believe that my (alleged!) brother is actually the product of a (revised) demon. Do I now have a defeater for R specified to my brother? Presumably, I would have such a defeater only if:

> The Analogous Claim$_3$: If P(R specified to someone else/revised demon hypothesis) is low or inscrutable, then one who believes the revised demon hypothesis about someone else has a defeater for R specified to that other person.

I think the Analogous Claim$_3$ is far from obviously true, keeping in mind that part of what we build into the *revised* demon hypothesis is the relevant *initial* warrant. For, after living with my brother for all those years while we were growing up, just seeing that his cognitive faculties are reliable, it seems that the reasonable moral of the revised demon hypothesis, applied to my brother, is that he beat the odds.

---

demon hypothesis] is true" ("Naturalism Defeated," 12). But suppose the evolutionary naturalist initially knows R specified to herself. She can then rule out both the evil demon hypothesis and the strong claim that the products of N&E are actually unreliable by way of something she knows (and so believes), that something being R itself. Plantinga is in an unenviable dialectical position if he must *first* claim that the evolutionary naturalist does not know R specified to herself in order to *then* argue that she has a defeater for R.

If you are not convinced, consider this. Suppose I have used a particular telephone, without incident, for years, and so am warranted in believing that phone is reliable. I then learn that phones made by company C—the company that made my phone—are almost all unreliable. Given my initial warrant in believing the phone's reliability, the obvious response to this news is that I lucked out. Given my noetic structure, involving initial warrant, this news constitutes *no defeater at all* for my belief in my phone's reliability. (It constitutes *no* defeater, not a defeated defeater. If it were a defeated defeater, there would have to be a defeater-defeater around. But there is not. For I still believe my phone was made by company C, and I still believe that P(my phone's being reliable/my phone's having been manufactured by company C) is low.)

As it goes with the phone, so I think it goes with my brother. And so I reject the Analogous Claim$_3$. (Another reason to reject the Analogous Claim$_3$ might be its similarity to the Analogous Claim$_1$, which one might already reject.) Once the Analogous Claim$_3$ is rejected, it seems like the Analogous Claim$_2$ ought to go too. And if we reject the Analogous Claim$_2$, then the revised evil demon case actually *undermines*, rather than supports, the Crucial Claim. Thus we have the makings of an argument *against* the Crucial Claim.

One might object that this argument against the Crucial Claim is wholly unpersuasive—or perhaps even almost question-begging—on the grounds that there is not enough epistemic distance between the denial of the Analogous Claim$_2$ and the denial of the Crucial Claim. One might object that, realistically, no one would deny the one who does not already deny the other. I think this objection has some force. But note that if this objection is a good one, it also shows that the attempt to *defend* the Crucial Claim by way of the Analogous Claim$_2$ is at best unpersuasive, at worst question-begging.

V

The cases that Plantinga actually offers for our consideration—the *standard* demon hypothesis, the *original* case of the hypothetical population—are, I think, clearly cases in which R (specified to oneself or to others) is defeated or unreasonable or jeopardized in some way. I have argued, however, that the cases Plantinga actually offers are not, in the form in which he originally presents them, epistemic analogues of the situation in which the devotee of N&E finds herself. And once the cases are tweaked into shape, so as to be epistemically analogous, they are no longer cases in which it is obvious that R—specified to others, oneself, whomever—has a defeater.

Moreover, the reason one would be inclined to doubt whether the revised demon hypothesis or the revised case of the hypothetical population involves a defeater for R is, I think, the *very same reason* one would doubt whether the evolutionary naturalist has a defeater for R. One sees that at the heart of each of the "revised" cases is an inference from low or inscrutable conditional

probability to defeat. Given that one knows that the Obviously False Claim is false, one suspects that such an inference—whether the inference involves the demonic or the Darwinian—might be mistaken. Also, the inferences in question might seem objectionably similar to the Crucial Claim, thus tainting an argument by way of these analogies with at least the suspicion that the question is being begged.

So I think Plantinga has not successfully defended the Crucial Claim. This does not mean that the Crucial Claim is false. Perhaps it is true. One response to my arguments above might be to *assert* the Crucial Claim, without argument, as a premise in the Evolutionary Argument. After all, Plantinga's original strategy of defending the Crucial Claim by way of analogous cases presupposes that there are *some* cases in which it is intuitively obvious that low or inscrutable conditional probability implies defeat. Perhaps, one might respond, the low or inscrutable P(R specified to oneself/N&E) is itself just such a case. But this response seems mistaken. For, as I noted earlier, I don't think that the Crucial Claim is intuitively obvious.

How could the Crucial Claim be defended by argument? One could develop cases in which it is obvious that low or inscrutable conditional probability implies defeat, and, moreover, clear that the evolutionary naturalist's belief in R specified to herself is epistemically analogous to the defeatee in those cases. I have argued that, so far, Plantinga has not successfully done this. In fact, I think this approach is probably a dead end. For once we build p's *initial warrant* into one's noetic structure, I cannot think of a single case in which it is intuitively obvious, or even intuitively plausible, that q is a defeater for p solely on the basis of a low or inscrutable P(p/q).

A second, and I think more promising, approach would be to defend a *general principle* about when p's being improbable or inscrutable on q implies that p is defeated by q, a principle that sanctions the Crucial Claim. This principle would presumably state what must be *added to* conditional improbability/inscrutability to get defeat. Plantinga has hinted at one ingredient of such a principle: "Among the crucially important facts, with respect to the question of the reliability of a group of cognitive faculties, are facts about their *origin*."[17] So perhaps an appropriate general principle will say that R's low or inscrutable probability on a claim relevantly related to the origin of one's faculties implies that that claim is a defeater for R.

As I said, I think the first approach is probably a dead end. At least, I can't think of any intuitively plausible cases of the relevant sort. And I confess utter failure at coming up with a plausible general principle that would sanction the Crucial Claim. So the bad news, for the Evolutionary Argument against Naturalism, is that I don't know how to help it along. But this news probably isn't so very bad. For the Evolutionary Argument's champion is surely more resourceful than I am and is likely to succeed where I have failed.

[17] "Naturalism Defeated," 11.

Plantinga, Naturalism, and Defeat

William Alston

I

Plantinga's "evolutionary argument against naturalism" has attracted much attention in the last few years. It received two simultaneous publications, in Radcliffe and White's *Faith in Theory and Practice* and in Plantinga's *Warrant and Proper Function*.[1] My references will be to the latter (hereinafter *WPF*). Having received many criticisms, Plantinga replies to them in a still unpublished paper, "Naturalism Defeated" (hereinafter ND). Despite all the critical fire, there are some serious problems with the argument that, so far as I know, have not been published to date. I begin with a quick sketch of the argument.[2] Further details will be supplied in the course of the discussion.

Plantinga first argues that the conditional probability of the (by and large) reliability of our cognitive faculties (R) on the conjunction of naturalism (N), current evolutionary theory (E), and the description of our faculties (C) is either low or inscrutable. He doesn't argue for this directly, but instead first argues for a like thesis with respect to "creatures a lot like ourselves on a planet similar to Earth. . . . Suppose . . . these creatures have arisen by way of the selection processes endorsed by contemporary evolutionary thought. . . . What is P(R/N&E&C), specified not to us, but to them?" (*WPF*, 223).

[1] Elizabeth S. Radcliffe and Carol J. White, eds., *Faith in Theory and Practice: Essays on Justifying Religious Belief* (Chicago: Open Court, 1993); Alvin Plantinga, *Warrant and Proper Function* (New York: Oxford University Press, 1993).

[2] I omit any consideration of what Plantinga calls "a preliminary argument against naturalism" (*WPF*, 228–229) in order to concentrate on what he calls the "main argument".

The argument proceeds by a consideration of five possibilities concerning the relations of beliefs and behavior for the population in question. They are:

1. No causal connection of belief and behavior.
2. Beliefs are the effects of behavior but are not among the causes of behavior.
3. Beliefs have causal impact on behavior but not by virtue of their content.
4. Belief is causally efficacious, in terms of content and otherwise, but is maladaptive.
5. Like 4., except that belief is adaptive.

In the case of all these except (5), Plantinga suggests that the probability of R on N&E&C, given that alternative, is low. Then, assuming that the alternatives are mutually exclusive and jointly exhaustive, he takes P(R/N&E&C) to be an average of that probability on each of the five alternatives, with those probabilities weighted for each alternative Ai in accordance with P(Ai/N&E&C). He then presents reasons for supposing that this weighted average will be "relatively low, somewhat less than ½", or at least that it is "very unlikely" that the average is "very high".

Plantinga then confesses that his "estimates of the various probabilities involved in estimating P(R/N&E&C) were (of necessity) both extremely imprecise and also poorly grounded. You might reasonably hold, therefore, that the right course here is simple agnosticism: one just does not know (and has no good way of finding out) what P(R/N&E&C) might be."

Because of complexities in the presentation of the final stages of the argument in *WPF*, I will rely here on the crisper formulation in ND. There, after the argument that P(R/N&E)[3] for the hypothetical population is either low or inscrutable, he goes on to infer a parallel conclusion for *us*.

> . . . if this is the sensible attitude to take to P(R/N&E) specified to that hypothetical population, then it will also be the sensible attitude towards P(R/N&E) specified to us. We are relevantly like them in that *our* cognitive faculties have the same kind of origin and provenance as *theirs* are hypothesized to have. And the next step in the argument was to point out that each of these attitudes—the view that P(R/N&E) is low and the view that this probability is inscrutable—gives the naturalist-evolutionist a *defeater* for R. It gives him a reason to doubt it, a reason not to affirm it.

Thus one who accepts N&E&C has a defeater for R and, therefore, for every belief formed on the basis of our faculties, i.e., for all our beliefs, including N and E. N&E is *self-defeating*. He finally adds that since if naturalism is true, then so, in all probability is evolution, naturalism *simpliciter* is self-defeating.

[3] In ND Plantinga omits C from that on which he argues that the probability of R is low. I will henceforward follow him in this.

II

I have two major objections to this argument, laying far and away the most emphasis, and giving the most space, to the second. (1) Not all the probability assignments are adequately defended. (2) Even if P(R/N&E) were either low or inscrutable, it would not follow that N&E is a defeater for R, possibly on Plantinga's own conception of defeaters, on one understanding of what that is. At least, I will argue, this doesn't follow on a maximally epistemically significant conception of defeaters. The first step in laying all this out is to consider how Plantinga understands conditional probability. That determination is crucial for both of the above criticisms and also for the evaluation of the inference from the hypothetical population to ourselves.

In chapter 12 of *WPF* there is a footnote telling us how to understand the probability claims.

> We could think of this probability in two ways: as a conditional *epistemic* probability, or as a conditional *objective* probability. Either will serve for my argument, but I should think the better way to think of it would be as objective probability; for in this sort of context epistemic probability, presumably, should follow known (or conjectured)[4] objective probability. (*WPF*, 220 n. 7)

But how does Plantinga understand 'epistemic probability' and 'objective probability'? He presumably did not feel called upon to explain them at this point because he had addressed that issue in chapter 9 of *WPF*, "Epistemic Conditional Probability: The Sober Truth." To be sure, that chapter doesn't completely solve our problem. For one thing, Plantinga distinguishes there *within* epistemic probability between an "objective component" and a "normative component", rather than simply contrasting epistemic and objective probability. With respect to the normative component he says the following.

> In asking after the normative component of such a probability judgment [of P(h/e)], we are asking what someone of "sound understanding," someone whose rational faculties are functioning properly, would believe [vis-à-vis h] . . . given that evidence. (*WPF*, 163).

Moreover, although he holds that the objective component can always be construed as a logical relation, in cases where e does not entail h this requires that e includes other propositions, in addition to those on which the probability is explicitly conditional. But what additional propositions? Plantinga gives us at least a rough indication.

[4] Plantinga needs to qualify 'conjectured' with something like 'rationally', as he does with 'believed' in *WPF*, 162. For it would be a matter of luck if genuine objective probability should "follow" irrationally conjectured objective probability.

> . . . the objective probability in question is indeed a logical probability, but it isn't one conditional just on the evidence. It is also conditional on other propositions: such propositions, perhaps, as that *The future will relevantly resemble the past* (the world is not, for example, a grue world) and perhaps *Simpler theories are more like to be true than complex ones*. Alternatively, the relevant set of possible worlds is not just the worlds in which the evidence is true, but some narrower class of worlds, perhaps specifiable in part in terms of similarity to what we think the actual world is like. (*WPF*, 162)

I take the general drift of these suggestions to be that the additional propositions in that on which the probability of h on e is (objectively) conditional are restricted to very general ones having to do with basic structural features of the world, rather than any that would ordinarily be thought of as evidence for or against h, or any other relatively specific proposition.

Returning to the presentation of the evolutionary argument against naturalism in chapter 12 of *WPF*, what I take to be the best way to align that presentation with the account of epistemic conditional probability in chapter 9 is to equate 'epistemic probability' in chapter 12 to 'the normative component' in chapter 9. As to Plantinga's statement in note 7 of chapter 12 of *WPF*, quoted above, that we should think of the conditional probabilities that figure in the epistemological argument against naturalism as objective rather than epistemic, I have no objection to that, provided we remember that in practice objective probabilities are decided by using the recipe quoted above from chapter 9 for determining epistemic probabilities. The rationale for this is that epistemic probabilities give us our best access to objective probabilities.

There is yet another large issue to be addressed concerning how to understand Plantinga's probability assessments. When asking about P(h/e) in the epistemic sense, are we asking what degree of credence one should give h on the basis of e alone, ignoring any further knowledge (beliefs) that are relevant to the epistemic status of h, or are we presupposing (the possibility of) other such items in the basis? I will term the first construal *parochial* and the second, *global*. Clearly we do make use of both construals. Suppose I ask, "If all you know about a woman is that her favorite novelist is Proust, what level of education would you suppose she has attained?" I am asking for the parochial probability of 'She has attained level n' on 'Her favorite novelist is Proust' (where the latter, of course, takes into account what Proust's novels are like). But when we make explicit probability judgments, they are almost always more global. Thus if we say something like "Given the meteorological conditions in the vicinity, the probability of rain in the next twenty-four hours is very high", we mean the basis to include not only the particular conditions in this vicinity now, but also some relevant body of general meteorological knowledge and hypotheses.

In Plantinga's argument we can find indications of both these construals. The parochial interpretation is suggested by the following.

1. The inference from the hypothetical population to us seems to presuppose this. In ND, we find Plantinga supporting it by pointing out that we are relevantly like that population "in that *our* cognitive faculties have the same kind of origin and provenance [according to N&E] as *theirs* are hypothesized to have" (11). A similar point is made in *WPF*, 229. But why should Plantinga think that this provenance is the only thing relevant to that probability? The obvious answer is that he is presupposing a parochial construal of P(X/Y) so that the support Y gives to X is the only thing relevant. Any additional information we have about us (and there will be a great deal of it) would make no difference.

2. The only thing Plantinga considers in arguing for evaluations of P(R/N&E) for the hypothetical population is the bearing of N&E on R.

3. The apparent reason for beginning the discussion with the hypothetical population is that it forces us to ignore other things we know about us. This would seem to be the force of the rationale, given in *WPF* (222), that he proceeds in this way "in order to avoid distractions". These "distractions" presumably consist of various other matters relevant to the epistemic status of R that we think we know about ourselves.

4. In ND he considers the "perspiration objection". It runs as follows. "The probability that the function of perspiration is to cool the body, given (just) N&E, is also low, as is the probability that Holland, Michigan, is 30 miles from Grand Rapids, given N&E. But surely it would be absurd to claim that these facts give the partisan of N&E a defeater for those beliefs" (13). Now it is noteworthy that Plantinga does not counter this objection by denying those probability assignments, as he would if global probability were under discussion. Clearly we have a great deal of evidence for both the above h's, but that does not figure at all in Plantinga's response. Instead his point is that a very low probability of h on e is not sufficient to make e a defeater for h. Rather than saying "On my conception of probability, 'Holland, Michigan, is 30 miles from Grand Rapids' is not improbable on N&E", he admits the improbability but denies that it makes the latter a defeater for the former. Indeed, he gives another example—that "You own an old Nissan" is improbable on "You own a Japanese car"—and denies that this improbability makes the latter a defeater of the former. And there are other examples that he treats in the same way. All this makes sense only if P(h/e) is being treated parochially.

So far as I can see, what indications there are on the other side are minuscule by comparison. The clearest example I can find is the following. In *WPF*, when presenting the extension to us with respect to an inscrutable probability, Plantinga writes: "Well, *if we have no further information*, then wouldn't the right attitude here, just as with respect to that hypothetical population, be agnosticism, withholding belief?" (229; italics added). This seems to assume that other information *is* relevant to the probability assessment; otherwise there would be no point in the italicized qualification.

It might seem, and it did seem to me earlier, that the treatment of proba-

bility in *WPF*, chapter 9, is a more important indication of the global interpretation. In particular, there is the point, noted earlier, that in typical judgments of the objective probability of h on e, the logical probability relation being claimed is not one that holds just between h and e, but between h and e plus "certain other propositions". And since, as noted, Plantinga rules in chapter 12 that we should be thinking of objective probability when considering the probability of R and N&E, this would appear to militate against a local probability interpretation.

But things are not always what they seem. To find a way out of this puzzle I will further deepen the problem by noting an apparent conflict between the treatments of the "objective" and "normative" aspects of epistemic probability in chapter 9. In spelling out the latter, Plantinga says the following.

> The rough initial idea, then, is that the normative component of the conditional epistemic probability of A on B is the interval containing the degrees of belief a rational person could have in A, provided she believed B and was aware that she believed B, considered the evidential bearing of B on A, *had no other source of warrant for B or its denial*, and had no defeater for the warrant. (167; italics added)

The italicized phrase seems to place the "normative component" squarely on the parochial side of the contrast, whereas the account of the "objective component" cited above makes it explicit that what the probability of h is conditional on, when we speak of P(h/e), includes more than e, thus rejecting a parochial interpretation. How, then, can it be, as Plantinga also says in chapter 9, that "In the typical probability judgment . . . these two components *coincide*"? (162).

The solution is found in a closer reading of the italicized phrase in the account of the normative component. It is only "other sources of warrant for B or its denial" that are ruled out of court in the specification of what the rational person would be taking into account. There is no parallel bar to taking into account the very general assumptions that Plantinga is thinking of as affecting the objective component.

How does this affect our choice between parochial and global construals for the argument in chapter 12? Clearly, it indicates that those alternatives were too unqualified. What we have just seen is that the probability under discussion in chapter 12 can be understood as parochial in a qualified sense. It is limited to the evidence specified, so far as *what counts specifically for or against h or its denial* is concerned. But that leaves open the relevance of very general considerations not included in e, as in the explanation of the "objective component". We could term this either a qualified parochial or a qualified global interpretation. But since it is considerations that count for or against h that are more salient in the local-global contrast, I come down on the side of a *qualified parochial* interpretation.

III

I now turn to my first main criticism of the argument, viz., that the probability assignments Plantinga makes in arguing for the thesis that P(R/N&E) is low, specified to the hypothetical population, are undersupported. Recall that Plantinga proceeds by distinguishing different alternatives for the relation of belief and behavior, alternatives he takes to be mutually exclusive and jointly exhaustive. He then takes P(R/N&E) to be "the weighted average of the probabilities for R on each of the possibilities—weighted by the probabilities of those possibilities". And given his estimates of both these sets of probabilities he takes P(R/N&E) to be "relatively low, somewhat less than one-half".

To evaluate Plantinga's reasoning, then, we have to consider both P(R/N&E&Pi), for each alternative Pi, and P(Pi/N&E) for each Pi. I begin with the latter, the probability of each alternative on N&E, since it can be quickly dispatched.

Plantinga has little to say about this. In *WPF* he says nothing beyond asserting of each P(P1/N&E) that "a substantial share of probability must be reserved for this option" or that it has "some probability". That doesn't help us in weighting the alternatives. In ND he only says that "the probabilities of P1 (epiphenomenalism) and P2 (beliefs causally affecting behavior but not in terms of their content) would be fairly high, given naturalism", and that the probabilities of the four alternatives[5] cannot sum to more than 1", thereby implying that the probabilities of the third and fourth must be fairly low. When we consider the task at hand, it is not surprising that Plantinga would say so little. Indeed, he has said too much. We know nothing about the population under consideration except that they are "*rational*: that is, they form beliefs, reason, change beliefs, and the like" and that "their cognitive systems have evolved by way of the mechanisms to which contemporary evolutionary theory direct our attention, unguided by the hand of God or of anyone else" (ND, 6). With no more than that to go on, how could we possibly make even a rough estimate of the likelihood that their beliefs do or do not have a causal influence on behavior, or that their beliefs are adaptive or maladaptive? How could we reasonably suppose such a likelihood to be high or low? If we could reason by analogy with what we know about ourselves, we might arrive at something, perhaps something that is pretty much the opposite of Plantinga's suggestions, for example, that the likelihood of causal efficacy of beliefs on behavior is much higher than epiphenomenalism. But the rules of the game forbid that. I see no alternative to throwing up our hands and declaring these probabilities *inscrutable*. But then the project of weighting the P(R/N&E&Pi) for each Pi by the P(Pi/N&E) breaks down.

[5] In ND the five alternatives of *WPF* are reduced to four, by omitting (1)—no causal connections of belief and behavior.

Suppose we ignore the weighting and simply take the average of the probabilities of R/N&E on each of the four alternatives. Though I have my doubts about some of those estimates, I am not at all clear about the matter and I will pass them by, so as to consider what follows if those estimates are accepted. Will that enable us to reach an objective result as to the overall value of R/N&E? A simple averaging will do the job only if we are justified in supposing the alternatives to be equiprobable as well as mutually exclusive. But why should we suppose this? We could invoke a principle of indifference if we were forced to make some assumption about their relative probability on N&E. But who or what is forcing us? In any event, a reliance on a principle of indifference on that basis could hardly be supposed to have the sweeping conclusion Plantinga draws from his argument.

Plantinga should be only mildly disturbed by this last criticism. For, as we saw above, after assessing P(R/N&E) for each of the alternatives he acknowledges that they are "both imprecise and poorly grounded", and hence "You might hold that the right course here is simple agnosticism" (ND, 11).

So the question is this. If we are unable to arrive at a confident estimate of P(R/N&E), and we assume N&E, does that imply that we "have good reason for being agnostic about R as well [assuming N&E]" (*WPF*, 229)? That does seem a sound judgment, at least in this sense. If all we know about the population in question is that they are cognitive agents, and N&E holds of them, and we are unable to say anything about the probability of R on N&E, then we are in no position to either assert or deny R of them. But, of course, what we are really interested in is R vis-à-vis us. And, it would seem, we are not subject to the above restriction. Far from it. We know a great deal about ourselves. And so the fact that we can't assess the probability of R on N&E alone will be of little interest if there are other things we know about ourselves that give R strong support. Suppose I cannot determine the probability that my wife is trustworthy on the evidence that she makes a mean Osso Buco Milanese. That will be of little moment if I know lots of other things that strongly support the proposition that she is trustworthy. So the verdict on the claim of agnosticism about R for us awaits the discussion below as to whether we do have considerable information about us that supports R.

IV

Now I come to my most serious criticism, which concerns the interpretation and significance of the claim that N&E is a defeater for R. Even if P(R/N&E) is low or inscrutable for us, how is that related to the claim that N&E is a defeater for R, and how is that latter claim related to the result to which all this is supposed to be leading, viz., that it is irrational for one who holds N&E to hold R, and hence to hold N&E itself. We will find that the attempt to deal with these questions takes us along a tortuous route. I shall be

spending most of my time on the "low probability" side of the disjunction of this with "inscrutable probability". The latter disjunct will receive brief treatment at the end.

First of all, we must be clear that Plantinga does not suppose that the probability of A's being low on B is sufficient, in general, to show that B is a defeater for A. This comes out most explicitly in the discussion of two of the objections to his argument in ND—the "perspiration objection" and the "Austere Theism objection". Briefly, the first of these objections points out that although "the probability that the function of perspiration is to cool the body, given (just) N&E is low, as is the probability that Holland, Michigan, is 30 miles from Grand Rapids, given N&E . . . surely it would be absurd to claim that these facts give the partisan of N&E a defeater for those beliefs" (ND, 13). Plantinga responds by denying the principle that a low or inscrutable probability of X on Y is a sufficient condition for Y's being a defeater of X. That, of course raises the question of what makes the difference between those cases, like the above, in which the low or inscrutable probability does not engender a defeater and those in which it does. I don't find that his response to these objections in ND does a very good job of answering that question. But in the presentation of his original argument, both in ND and, at somewhat greater length in *WPF*, he presents some features of the N&E–R relationship that can be seen to differentiate it from the "perspiration objection". The crucial claim is that N&E involves a claim about the *origin* of our cognitive faculties. And because of this, if there is a low, or inscrutable, probability, given such an origin, that those faculties are reliable, then this does provide a defeater for R, the claim that they are reliable. Whereas N&E is not related in that way to the claims about perspiration or the distance between Grand Rapids, Michigan, and Holland, Michigan.[6]

This point is spelled out by presenting some analogous cases in which the origin of a belief, or of beliefs of a certain type, is such as to make it improbable that the belief(s) would be true, cases in which, Plantinga claims, it is obvious that this provides a defeater for the supposition that beliefs with such an origin are generally true.[7]

> Suppose I believe that I have been created by an evil Cartesian demon who takes delight in fashioning creatures who have mainly false beliefs (but think of themselves as paradigms of cognitive excellence): then I have a defeater for my natural belief that my faculties are reliable. Turn instead to the contemporary version of this scenario, and suppose I come to believe that I have been captured by Alpha-Centaurian superscientists who have made me the subject of a cognitive experi-

[6] In correspondence Plantinga has pointed out some other differences. We have strong independent evidence for both of the claims in the objection, but no strong (or any) evidence, in the context of the argument, that is not question begging, by reason of assuming the very point at issue, viz., that our cognitive faculties are generally reliable.

[7] I am leaving the inscrutability alternative for later treatment.

ment in which the subject is given mostly false beliefs: then, again, I have a defeater for R. (ND, 11)

> Compare the case of a believer in God, who, perhaps through an injudicious reading of Freud, comes to think that religious belief generally and theistic belief in particular is almost always produced by wish fulfillment. Such beliefs, she now thinks, are not produced by cognitive faculties functioning properly in a congenial environment according to a design plan successfully aimed at truth; instead they are produced by wish fulfillment, which, while indeed it has a function, does not have the function of producing true beliefs. Suppose she considers the objective probability that wish fulfillment, as a belief-producing mechanism, is reliable. She might quite properly estimate this probability as relatively low. . . . But then . . . she has a defeater for any belief she takes to be produced by the mechanism in question. (*WPF*, 229–230)

In both *WPF* and ND Plantinga goes on to claim that one who believes that his cognitive faculties are produced according to N&E, and who recognizes that the probability of those faculties being generally reliable, given that provenance, is low, thereby has a defeater for the belief that those faculties are reliable, and, by derivation, for any belief produced by those faculties, including N&E.[8]

This looks to be a promising line of argument, but it needs further spelling out. Just what origin does N&E take our cognitive faculties, in particular our belief-forming faculties, to have? The impression given by Plantinga's exposition is that N&E implies that these faculties result from evolutionary mechanisms, in particular natural selection. For one thing, the argument for a low value of P(R/N&E) is developed in opposition to the position of Popper and Quine that this origin makes it likely that these faculties are reliable. Moreover, in *WPF* Plantinga comes close to an explicit formulation of this reading. "Suppose we think N&E is true: we ourselves have evolved according to the mechanisms suggested by contemporary evolutionary theory, unguided and unorchestrated by God or anyone else" (229).[9]

But things are not always what they seem. Remember the five possibilities for the relation of belief and behavior distinguished by Plantinga. On only the fifth of these does belief causally affect behavior, by virtue of its content, in an adaptive fashion. On the other four either belief does not causally affect behavior at all (1 and 2), or it does but not by virtue of its content (3), or it does by virtue of its content but not so as to render behavior adaptive (4). It is only on the fifth possibility, on which (true) belief influences behavior so as to render it likely to be adaptive, that natural selection would favor reliable be-

[8] In *WPF* Plantinga also cites the case of an "undermining" defeater (one that is not even probabilistically incompatible with the target belief but that indicates that S's ground for that belief lacks force). But this is much less analogous to the N&E–R relationship. Plantinga does not present N&E as undermining the evolutionary naturalist's reason for accepting R. Indeed, he says nothing as to what such a reason might be.

[9] The first conjunct is supposed to come from E and the second from N.

lief-forming faculties. On the other four, as Plantinga says with respect to the first two, "they are invisible to evolution" (*WPF*, 223). But in that case, evolutionary mechanisms plays no role in their genesis—in our acquiring belief-forming faculties at all, or in our acquiring such faculties of one kind rather than another. And so, unless we can limit the relevant possibilities to (5) (and Plantinga's argument essentially depends on not doing so), we have to say that N&E can't be or imply a thesis about the origin of these faculties in that way. "The mechanisms suggested by contemporary evolutionary theory", in particular natural selection in forms of various degrees of sophistication, can have nothing to do with it.

But all is not lost for Plantinga's argument. The N component of N&E may come to the rescue. For, on alternatives 1–4, N guarantees that our acquisition of these faculties is "unguided and unorchestrated by God or anyone else". So if their origin cannot be explained by evolutionary mechanisms, N&E is left with nothing to say about the origin that would serve to explain why we have cognitive faculties of the sort we do. In particular, N&E, on any of the first four alternatives, would have nothing to say that is relevant to the question of their reliability. It is conceivable that natural laws should be discovered such that the possession of whatever psychic factors influence behavior in an adaptive fashion would carry with it, by nomological necessity, the possession of beliefs with propositional content that themselves have no such influence. But even so, this would provide no answer to the likelihood of the reliability of the belief-forming faculties. So far as the question of reliability is concerned, we may as well say that the faculties arose by sheer chance, by some sort of random drift. And this is just as good a reason for holding that on (1) to (4) the probability of our belief-forming faculties being reliable cannot be higher than ½, and would seem to be significantly lower.

Thus Plantinga has a strong case for the thesis that *if* there is a low probability of R on N&E, that has a significant bearing on the rationality of holding both N&E and R. They are in such conflict that it would be irrational to hold both. But to evaluate Plantinga's full argument and the conclusion thereof, we need to ask whether this implies that one who holds N&E has a defeater for R (rather than vice versa), and if she does whether this makes it irrational for her to hold R (rather than irrational for her to hold N&E). And the first step in looking into that is looking at the concept of *defeat* being employed. To this I now turn.

V

In both ND and chapter 11 of *Warranted Christian Belief* (hereinafter *WCB*),[10] entitled "Defeaters and Defeat", there is extensive discussion of the issue. In *WCB* we find the following definition.

[10] New York: Oxford Univesity Press, 2000.

(*D**) *D* is a purely epistemic defeater of *B* for *S* at *t* iff (1) *S*'s noetic structure *N* at *t* includes *B* and *S* comes to believe *D* at *t*, and (2) any person *S** (a) whose cognitive faculties are functioning properly in the relevant respects, (b) who is such that the bit of the design plan governing the sustaining of *B* in her noetic structure is successfully aimed at truth (i.e., at the maximization of true belief and minimization of false belief) and nothing more, (c) whose noetic structure is *N* and includes *B*, and (d) who comes to believe *D* but nothing else independent of or stronger than *D*, would withhold *B* (or believe it less strongly). (*WCB*, 363)

Plantinga's reason for concentrating on the notion of a '*purely epistemic* defeater' is that he wants to identify the kind of conflict between D and B that would make it a violation of the aim at truth to hold them both, rather than a violation of the aim at, for example, survival or psychological comfort. Even if believing B along with D would make one extremely uncomfortable, that would not (tend to) make D a purely epistemic defeater of B.

In this chapter of *WCB* Plantinga also makes a couple of further distinctions that are relevant to understanding the notion of a defeater he is using. First the kind of defeater picked out by D* is a *rationality defeater*. "Given belief in the defeating proposition, you can retain belief in the defeated proposition only at the cost of irrationality" (*WCB*, 359). This is distinguished from a *warrant* defeater that prevents the defeated proposition from enjoying enough warrant to count as known (even if true), but is not such as to prevent one from rationally accepting the proposition. Plantinga's example is the familiar barn case. S, on driving through the countryside sees something that looks like a barn, and he forms the belief that it is a barn. And so it is. But, as it happens, the area in question contains numerous barn facades that would look just like a barn to S, given his distance, speed, angle of vision, etc. S is rational in taking what he sees to be a barn, but since it is just a matter of luck, given what is in the neighborhood, that what he saw at that point was a barn rather than a barn facade, he can't be said to *know* that it was a barn. Note that a rationality defeater is a *belief*, something that belongs to the subject's noetic structure, whereas a warrant defeater can be some fact external to that structure.

The other distinction is between *internal* and *external* rationality. "*Internal* rationality is a matter of proper function 'downstream from experience'. . . . *External* rationality, by contrast, is a matter of the proper function of the *sources* of experience . . ." (*WCB*, 365). Spelling this out a bit, I take it that internal rationality has to do with the internal economy of one's noetic structure—its coherence—while external rationality has to do with the way in which new items are added to that structure. Plantinga's example of the latter in this passage has to do with the role of pride in distorting one's evaluations of one's own accomplishments. In a moment I will explore the way this distinction bears on Plantinga's use of 'defeater', something that I find not entirely clear from the texts at my disposal.

To return to the definition, D*, one thing on which we obviously need more light is the stipulation that the hypothetical subject's "faculties are functioning properly *in the relevant respects*" (italics added). How is the italicized phrase to be fleshed out? Plantinga does not tell us. A natural reading would be that these respects include anything that is relevant to the epistemic status of B and D and to their logical (including probabilistic) relationships, and hence that is relevant to what would be the most rational response to their incompatibility. This would at least include the kind and degree of incompatibility relation between B and D, what intrinsic positive epistemic status is enjoyed by each along with whatever support each receives from other components of S's noetic structure.[11] But why wouldn't the degree of warrant enjoyed by D and B also be a "relevant respect", even where determinants of degree of warrant lie, at least in part, outside S's noetic structure? Wouldn't that be highly relevant to determining whether it would be more rational to reject B or to reject D? Isn't it clear, from Plantinga's epistemological perspective, that in case of incompatibility it is most rational to stick with the more warranted contender? Here the question of whether it is internal or external rationality or both (or alternatively, the question of what kind of proper function) is involved in D* is crucial. If degree of warrant is a relevant respect, then the rationality defined by what one whose cognitive faculties are functioning properly *in the relevant respects* would do would include external, as well as internal, rationality. But it is clear that Plantinga is not prepared to be that inclusive. This comes out most clearly in his insistence, in both ND and *WCB*, that a belief with no warrant can be a defeater. ". . . a belief D can defeat another belief B, for me, even if D has little or no warrant for me" (ND, 24). "Suppose I hold a belief *B*, but then come to accept a belief *D* that goes against *B* in some way, but this belief *D* I accept has no warrant. Can it still be a defeater for *B*? I should think so" (*WCB*, 364). In these passages Plantinga does not say that B does have warrant, but since he gives no hint that a greater warrant for B would prevent D's from defeating B, I shall take it that he means to be committing himself to the thesis that a belief with a lesser (down to no) warrant can defeat a belief with greater warrant. This does clearly indicate that the relative warrant of D and B is not included in the "relevant respects" specified in D*. But, furthermore, Plantinga says "it is possible for a belief that is irrationally acquired to be a defeater, even for a belief that is rationally acquired" (*WCB*, 363). Here is one of his examples.

> Suppose I've always thought you a genial sort who is rather well disposed to me. Unhappily, I start sinking into a paranoid condition; because of cognitive malfunction, it comes to seem to me that you are, in fact, trying to harm me by destroying my academic reputation. Because of the cognitive malfunction, this just

[11] If the rationality involved in D* is restricted to internal rationality, then it is only support from within the noetic structure that is relevant. Otherwise support from outside is or might be relevant. I will return to this issue.

> seems wholly obvious to me. . . . Can my belief D that you are trying to destroy my reputation serve as a defeater for my belief B that you are favorably disposed toward me? (*WCB*, 365)

He answers this in the affirmative and supports this answer by invoking the distinction between internal and external rationality. In terms of that distinction, he says that the belief that you are out to get him is *externally* irrational "because it arises from sources of doxastic experience that are not functioning properly". (By 'doxastic experience' Plantinga refers to the *sense* of truthfulness or falsity that accompanies beliefs.)

> By virtue of their malfunction, however, my experience is such that I am powerfully impelled to believe D, that you trying to ruin me. This now seems to me much more obvious than that you are favorably disposed toward me: the doxastic evidence for D is much stronger than that for B. What internal rationality calls for, under those circumstances, therefore, is my giving up B; I have a defeater for it in D, even though D is arrived at irrationally. I can therefore have a defeater D for a belief B, even where B is rationally held and D is irrationally acquired. (*WCB*, 365)

Thus, according to Plantinga, a less warranted belief, D, can defeat a more warranted belief, B, and an (externally) irrationally acquired belief can defeat an (externally) rationally acquired belief, provided that *internal* rationality in each case calls for retaining D rather than B. Note that this example, and other examples he gives, supports my conjecture that internal rationality concerns the internal economy of one's noetic structure, what one can access concerning it "from the inside". What it leaves out, inter alia, is the relation of that structure to outside factors, including, crucially, facts concerning the origin of its components, *at least so far as these facts do not concern relations of the new component to already existing components*. Thus, to a first approximation, we may think of the "relevant respects" mentioned in D* as restricted to what it takes to maximize internal rationality.

Now we must examine another clause of D*, the subclause of (2) that requires of the hypothetical subject, whose potential doxastic reaction to the D-B conflict determines whether D defeats B, that it be one "(c) whose noetic structure is *N* and includes *B*". We will see that this clause reinforces the above conclusion that the "relevant respects" include those that are concerned with internal rationality but not those concerned with external rationality. Of course, one rationale for the clause is that it ensures that the model subject has the qualifications for the task, in particular has the right noetic structure. But a close examination of the way Plantinga handles certain examples suggests that it plays the further role of making the point that positive and negative probability relations among components of the doxastic struc-

ture are taken into account by our ideal subject in deciding how to resolve the D-B conflict.

> I have thought for some time that you once spent a year in Aberdeen, Scotland; you tell me (soberly, with no hint of teasing or joking) that you have never been to Scotland, although you once planned to go but were prevented at the last moment. I thus learn that you have never been to Scotland; the noetic structure that results from adding this belief to the noetic structure I have at *t* is irrational. . . . What rationality requires is that I change belief. . . . But not just any old change. . . . What needs to be given up is my belief that you once visited Aberdeen. . . . You might think that other changes would be . . . consistent with rationality. . . . Perhaps I could give up the belief that you are now truthful, or the belief that you are mentally competent, or the belief that you are capable of distinguishing Aberdeen, Scotland, from Aberdeen, South Dakota. *And perhaps these changes would be rational with respect to some noetic structures—ones, perhaps, in which I have enormously powerful evidence for your having been in Aberdeen . . . or structures in which I have good reason to doubt that you are telling the truth. But these are quite different from the structure I do in fact display at t, which involves my being quite properly sure that you are telling the truth, and also involves my having little more by way of support for the belief that you have been to Scotland, than a sort of vague memory to the effect that I once learned this. With respect to this noetic structure, these changes would not be rational. That is why the defeater is a defeater for my belief that you have been in Aberdeen, rather than for some other belief.* (ND, 33–34; italics added)

The italicized portion makes it crystal clear that Plantinga is prepared to take into account the relevant epistemic status of D and B in deciding whether D defeats B rather than vice versa, or rather than D defeating something else if anything. Is this incompatible with the restriction to *internal* rationality? Perhaps not, so long as the only sort of epistemic statuses in the picture are those that consist in the logical and probabilistic relations of the beliefs in question to other beliefs in the structure. And what is said here is compatible with that (though Plantinga's speaking of being *properly* sure that you are telling the truth evokes echoes of his concept of warrant, which is by no means confined to internal relations within the noetic structure).

But this example, and other examples, do raise one question about the understanding of internal rationality that I have not yet addressed. When Plantinga speaks of "powerful evidence for your having been in Aberdeen" and "good reason to doubt that you are telling the truth" and "little more by way of support", are these to be understood as referring to objective relations of *powerful evidence for* and *good reason to*, which obtain, where they do, regardless of whether S believes they do or is justified (warranted) in believing they do? Or are they to be understood as referring to what S believes about such relations (with or without justification)? Or are both required?[12] My previous in-

[12] We shall see later that Plantinga holds that a low probability of B on D does not render D a defeater of B unless S believes that this low probability relation obtains.

terpretation of internal rationality as concerned with the "internal economy" of the noetic structure is ambiguous between these interpretations. It could be "internal" by virtue of the components of the economy without having to include, among the components, the relations between the components that make the economy what it is. Furthermore, if we were to insist that beliefs in the logical and probability relations that make the economy what it is be among the components of the structure, we would be off on an infinite regress. For having included beliefs in these relations among the components, we would then need more relations between these new components and the others. And then, having added those relations . . . We would never get the complete economy specified. For this reason, and since a natural reading of epistemic terms like 'powerful evidence' would be in terms of objective relationships, and in the absence of any statement by Plantinga known to me on the point, I will assume that a reading in terms of objective evidential relations is compatible with the restriction to internal rationality.[13] That leaves the way in which Plantinga handles this and other examples compatible with the ignoring of the relative *warrant* of D and B in considering what a subject whose cognitive faculties are functioning properly *in the relevant respects* would do about the D-B conflict. For in Plantinga's epistemology, warrant generally depends on more than evidential relations within a noetic structure. Moreover, it leaves the treatment of the examples compatible with allowing that an irrationally acquired belief can be a defeater of a rationally acquired defeater, provided what makes the former irrational is, at least in part, something outside the noetic structure.

But there is another consideration that seems to disturb this picture of sweet harmony. In ND, after saying that "a belief is rational in a certain set of circumstances when it is a healthy or sane belief to hold in those circumstances" (21), he continues:

> The relevant circumstances have a two-tiered character. First, there is my *noetic structure;* an assemblage of beliefs and experiences (and other cognitive states such as doubts, fears and the like) together with various salient properties of these states and relevant relations obtaining among them. Let's oversimplify and think just of beliefs and experiences. A description of a noetic structure would include a description of the strength of each belief, of the logical relations between that beliefs and others, and of the circumstances (crucially including experiences) under which the belief in question was formed and sustained. Not all beliefs are formed in response to experience (together with previous belief), and it may be that some beliefs are formed in response to experience, previous belief and still other circumstances: let's use the ugly but popular term *doxastic input* to denote

[13] We should also note that in *WPF* the claim that N&E is a defeater for R, for us, is sometimes qualified by provisos like "If we have no further information [about R]" (231), or "if we have no independent evidence [for R]" (233). But unless the epistemic status of R vis-à-vis the rest of S's noetic structure were relevant to the claim that N&E is a defeater for R, these provisos would have no point.

> whatever it is that beliefs are formed in response to. Then we can say that a description of S's noetic structure would include an account of the doxastic input to which S has been subject, as well as an account of the doxastic responses thereto. (ND, 21–22)

This account obviously is in line with my earlier supposition that the relations that structure a noetic structure can do so whether or not the subject believes them to obtain. In speaking of relations obtaining between beliefs, Plantinga gives no hint that he thinks such relations are (must be) believed by the subject to obtain. But there is a further feature of the account that does not jibe with the developing picture. According to the passage just quoted, "a description of a noetic structure would include a description . . . of the circumstances (crucially including experiences) under which the belief in question was formed and sustained". And it is made explicit that these circumstances are not restricted to other beliefs. Hence it would seem that a full description of a noetic structure would include a full account of the "circumstances" under which the belief is formed and sustained. And that, in turn means that the description would include the degree of warrant of each belief, since in Plantinga's epistemology warrant is a function of relevant facts about the forming and sustenance of the belief in question. But then, given that relevant facts about S's noetic structure are among those that are relevant for the response of our paradigmatic subject (in D*) to the D-B conflict, how can the relevant degree of warrant fail to be included in those considerations? And hence how can it be rational for our ideal subject to prefer a less warranted to a more warranted belief? And, to put it in terms of rationality, how, given the relevance of the noetic structure as so described, can the ideal subject fail to make the response that is most rational, externally as well as internally?

We seem to have a flat incompatibility between these texts. For a while I thought that I was missing something in the account of noetic structures just quoted, something that neutralized the apparent implication that the full description of a noetic structure would include a specification of the degree of warrant enjoyed by each constituent belief—though I had not been able to find any such something. But in a recent correspondence Plantinga writes in response to a question I posed, "As for a noetic structure, I was thinking of it as including something like an index of degree of warrant for each belief". And so the incoherence won't go away. For the moment I will simply live with two versions of 'defeat'; in one, relative degree of warrant is relevant to what defeats what, and in the other it is not relevant. The first version takes the noetic structure to include everything specified in the above quotation, including degree of warrant for each belief, and takes all that as relevant to the hypothetical subject's response to the D-B conflict. Whereas the second version places a constraint of only internal rationality on the response of the hypothetical subject, where internal rationality does not require taking into account the "external" features of how the belief is formed and sustained. In this

chapter I will go with the second, leaner version, which is more strongly supported by the texts.

VI

So much, for the moment, for Plantinga's notion of a defeater. Now we need to relate the claim that N&E is a defeater for R to the conclusion that Plantinga is primarily concerned to establish—namely, that one who holds N&E cannot rationally hold R (and, as a further conclusion, cannot rationally hold any belief at all, including N&E). Though Plantinga does not sufficiently underline the point in the published statement of the argument in *WPF* (and in its predecessor in Radcliffe and White's *Faith in Theory and Practice*), N&E's being a defeater for R is not by itself sufficient for the further conclusion. For one thing a defeater can itself be defeated (*WPF*, 233; ND, 12). Defeaters, including defeaters for defeaters, can be divided into *rebutters* and *undercutters*. The former is what is defined by D*. An undercutting defeater is a belief that indicates not that the defeatee is false but that one's grounds for it are inadequate or neutralized. Since a defeater is a belief, it might, like any belief, be itself defeated in either of these ways. If it is, its efficacy as a defeater is thereby nullified.

Plantinga has another way of allowing that N&E's being a defeater for R may be insufficient to render it irrational for S (who accepts N&E) to accept R. S might have such strong support for R as to outweigh the negative impact of N&E (*WPF*, 233). Plantinga is sufficiently alive to the necessity of both these conditions to devote several pages of *WPF* to arguing that one can't have non-question-begging support for R, and similarly that any supposition that N&E is itself defeated is also question begging, on the grounds that any such suppositions presuppose R, that is, assume that our faculties are generally reliable.

By virtue of treating *no defeater defeaters* and *no sufficient independent evidence for R* as requirements that are additional to N&E's being a defeater for R, Plantinga runs into further internal difficulties. For note that where such requirements are not satisfied, it is by virtue of features of S's noetic structure. And, as we have seen, D* implies that features of the noetic structure are relevant to whether our ideal subject would resolve the D-B conflict by rejecting B, i.e., are relevant to whether D *is* a defeater for B. So how can they be requirements additional to D's being a defeater for B? I don't take this to be nearly as serious a problem for Plantinga as the earlier problem as to whether relative degree of warrant counts as to what the most rational resolution of the conflict would be. It is almost just a matter of bookkeeping. If they are extra requirements, then the definition of 'defeater' will have to be modified in such a way that they are not requirements for D's being a defeater for B. Otherwise, we can just bill them as extra requirements.

But there is a bit more than convenience of bookkeeping involved in this choice, which gives us reason to prefer the alternative of taking these requirements to be necessary for defeater status, as D* (on my reading) implies. There are three advantages of this choice.

1. As just noted, this permits the retention of D* as a definition of 'defeater'.

2. Furthermore, it fits better the intuitive notion of a defeater. Plantinga formulates this in several places.

> . . . the basic but rough answer is that defeaters are reasons for changing one's beliefs in a certain way. (ND, 20)

> . . . the basic idea is that when S acquires a defeater for B, she acquires a reason for modifying her noetic structure in a certain way. (ND, 31)
> Acquiring a defeater for a belief puts you in a position in which you can't rationally continue to hold the belief. (*WCB*, 359)

The first two formulations are very unspecific. Modify beliefs in *what* way? And how strong a reason? But the third formulation ties things down. The change is a deletion of the belief for which the defeater is a defeater. And the reason is strong enough to make it irrational to continue to hold that belief. To fit that specification we need the "extra" requirements.

Apart from Plantinga's delineation of the intuitive idea, it must be noted that 'defeat' is a term of considerable epistemic strength. It seems unintuitive to call D a defeater for B if its defeating efficacy is overridden by a higher-level defeater or by a stronger support for B. In that case, we would have what might more felicitously be termed a '*prima facie* defeater' or a 'candidate defeater'. We need the extra requirements to get the full epistemic force the term 'defeater' suggests.

3. Reserving 'defeater' for what satisfies the "extra" requirements enables us to avoid an uncomfortable shift if one or another of these requirements are not satisfied. Suppose being undefeated is not required for D's being a defeater of B. Then it turns out that D is itself defeated. Is it still a defeater of B, albeit a defeated one, or does it lose that status? If the defeater of D is added to the noetic structure after D, we can think of that addition as changing D's status from being a defeater to not being a defeater. But what if the defeater of D was there all along, but unnoticed? Or it was noticed but its defeating of D was not? In that case, did D defeat B prior to the noticing? We have a similar problem with strong support for B that turns up or comes to be realized at a later point. Building the "extra" requirements into what it takes for a belief to be a defeater, as in D* on my reading, will avoid these problems.

It is time for a review of my interpretation of Plantinga's account of defeaters. First, I will ignore indications that Plantinga does take relative warrant of D and B (and perhaps other "external" factors as well) to be relevant to

the question of whether D defeats B, except where explicitly noted to the contrary. I will be working with the reading of D* that restricts "relevant respects" and features of the noetic structure to matters that are relevant to "internal rationality". But, as noted, I will take the noetic structure to be organized by objective logical and probability relations between beliefs, relations that may or may not be noted by S. This means, inter alia, that positive and negative "support" for D and B from other beliefs and from experience are relevant to the reaction of our ideal subject to the D-B conflict, and hence to whether D is a defeater of B. Moreover, this support is to be evaluated not only by the number and content of supporting beliefs but also by the epistemic status of these beliefs, *insofar as it is a function of the relation of the belief to other beliefs and to experience*. Thus an (internally) rational response to a conflict is determined by the *internal economy* of the noetic structure, so understood. Note that this implies that although a defeat relation so understood is of undoubted epistemic significance (it does constitute a reason for abandoning a belief), its epistemic significance is limited, so long as we ignore the relative degrees of warrant of constituents of the noetic system and other external influences on epistemic status. I will return to this last point when I move to critical remarks on this account of defeat and on Plantinga's use of it in his evolutionary argument against naturalism.

VII

It is high time to move away from the general notion of defeat and look more specifically at Plantinga's claim that anyone who holds N&E thereby has a defeater for R (and, by derivation, for any belief whatever, including N&E). Remembering Plantinga's disjunction of two views as to the probabilistic relation between N&E and R, I will be working with the *low value of P(R/N&E)* alternative, adding some brief remarks on the *inscrutability of P(R/N&E)* alternative at the end of the chapter. Remember also that Plantinga argues that what he takes to be the extra requirements of *N&E's not itself being defeated* and of *the lack of (sufficient) independent evidence for R* are satisfied. Later I will suggest some doubts about the latter claim, but for now I will assume Plantinga is right in both cases, thus allowing him those parts of what it takes, on my preferred bookkeeping, for N&E to defeat R.

To be sure, if I were to concede to Plantinga in this case everything that D* requires for a defeater relation, I would pass on to the question of the significance of N&E's defeating R, in this sense of the term. But I am by no means prepared to do that. The main worry is this. For this stage of the argument I also concede a low value of P(R/N&E), despite the doubts to which I alluded in section 3. But the considerations that support that result equally support a low value for the reverse relation, P(N&E/R). For if the naturalistic evolutionary story of the development of human cognitive faculties makes it un-

likely that they are reliable, then, by the same token, it they are reliable that makes it unlikely that they arose in the naturalistic evolutionary way. If certain meteorological conditions make a thunderstorm unlikely, then if a thunderstorm occurs it is unlikely, in the absence of other relevant information, that those meteorological conditions prevailed. So the question is: Why take it that N&E defeats R, rather than that R defeats N &E? Or in terms of D*, why suppose that a subject working under the conditions specified in D* would withhold R, rather than withhold N&E? There could be other grounds for Plantinga's preference. R could be defeated by other parts of the noetic structure and N&E not. But Plantinga gives no hint of this, and let's assume that neither is otherwise defeated. Or one or the other could have much stronger independent support within the noetic structure than the other. As noted, Plantinga argues that it would be question begging in this context to suppose that R enjoys strong independent support. And though Plantinga does not discuss the issue in these texts, he does not believe that the N conjunct at least of N&E enjoys strong independent support. But rather than turn my argument into an ad hominem, let me just say that unless N&E enjoys significantly stronger independent support than R, there does not appear to be, within the terms set by Plantinga, any reason to prefer taking N&E to defeat R rather than taking R to defeat N&E.

VIII

But perhaps the game Plantinga is playing has been misconstrued. So far I have not cited passages like the following (italics added).

> A defeater for a belief *b*, then, is another belief *d* such that, given my noetic structure, I cannot rationally hold *b*, *given that I believe d*. (*WCB*, 361; italics added)

> . . . you have a defeater for one of your beliefs B just if you acquire another belief D such that, *given that you hold that belief*, the rational response is to reject B. (*WCB*, 366; italics added)

> . . . no noetic structure to which S can rationally move (*given that she accepts D*) will contain B. (ND, 33; italics added)

What these formulations suggest is that the question of what defeats what (in terms of D*, what the most rational response to the D-B conflict would be) is one that is raised in a certain context, a context that is defined by S's holding D. Since that is a presupposition of the question being raised, part of the content of the question itself, the rejection of D is not one of the options between which a choice is to be made. Applying this point to our central concern here, the question of what rationality requires is raised for the person

who accepts N&E. The question is as to what rationality requires for a person who continues to satisfy that description. Hence there is no problem as to what reasons there are for a preference between D's defeating B and B's defeating D. The latter alternative has been blocked out before we get started.

Perhaps that is the way Plantinga is thinking of the matter. But if so, this markedly affects the epistemic significance of the question he is raising and hence the epistemic significance of the answer to the question. The question is not without significance. It is sometimes relevant to consider what a person is committed to, on pain of irrationality, given that he accepts a certain proposition. This can be useful in persuading him to give up the proposition in question. Once he sees how the proposition would render his belief system seriously incoherent or inconsistent, he may give it up. Presumably something like this is involved in Plantinga's argument. Addressed to the proponent of N&E, it can be construed as an appeal to such a person to take seriously the dismal situation that confronts him if he continues to hold fast to N&E. Either he must deny that his cognitive faculties are reliable, in which case he is without adequate rational basis for any beliefs whatever, or he must suffer crucial incoherence in his belief system.

But how about defeat, as defined by D*? Does this "rationality relative to holding N&E fixed" interpretation enable us to conclude that for S (an advocate of N&E) N&E is a *defeater* of R in the sense defined by D*? Again, this depends on how "relevant respects" are specified, and on what is included in the description of S's noetic structure. What's the best we can do along this line? Well, suppose we go along with Plantinga in supposing that N&E itself is undefeated and that there is no significant non-question-begging support for R. (Remember that we have decided to include that in what is required for N&E's defeating R.) But now what about the hypothetical subject's cognitive faculties functioning proper in the *relevant respects*. I have already pointed out that there would seem to be no sensible alternative to allowing the relative epistemic status of N&E and R to be relevant to determining which, if either, defeats the other. How could their status not be relevant? Even if we restrict ourselves to the "internal rationality" alternative, we will at least have to count as relevant the intrinsic evidence of each (proper basicality) and the confirmation and disconfirmation each has from within the subject's noetic structure. But then it seems that we are still driven to the former conclusion. If N&E is no better supported internally than R, the situation still fails to tell in favor of N&E's defeating R, rather than the reverse, if each has a low conditional probability on the other. Otherwise put, if that is the case, then unless N&E is significantly better internally supported than R, we cannot conclude that a subject whose cognitive faculties are properly functioning in the relevant respects would reject R, rather than rejecting N&E.

By the way, I can easily imagine Plantinga's welcoming this conclusion. For it provides the basis for a strong argument against N&E. "Either N&E is woefully undersupported by evidence, or it leads to self-refuting conse-

quences. In either case, it is irrational to hold it."[14] If he would like to appropriate this in the next version of his argument, I make him a free gift of it. But my concern in this chapter is to critically evaluate the argument as he has presented it.

To continue, if my conclusions thus far are sound, the only way Plantinga can derive the conclusion that the ideal cognitive subject will reject R, assuming that N&E lacks significantly stronger internal support, is to "cook" the account of relevant respects in such a way as to reflect the "retain no matter what" status of N&E. That would amount to presenting the above argument as to what the internal coherence of one's belief system commits one to if one accepts N&E, in the guise of an argument for what an ideally rational subject would do about the N&E–R conflict in the light of all features of its noetic structure that are relevant to that resolution. But those clothes simply do not fit the "internal coherence of belief system with N&E fixed" argument. They are quite different arguments, quite different trains of reasoning, suitable for different purposes, and it is false advertising to present the former as if it were the latter. Whereas the former is suited to convincing (or trying to convince) an N&Eer of the error of his ways, the latter is suited to consider how a subject, ideally rational in the relevant respects would respond to a probabilistic inconsistency between N&E and R. To repeat, if we give a sensible reading of "relevant respects", we simply can't conclude that N&E defeats R (in the sense of 'defeat' given by D*) unless N&E is significantly better supported internally than R.

Here is another way of seeing this. Consider a favorite example of Plantinga's, Descartes's reference to a madman, M, who thinks his head is made of glass. Plantinga uses the case of M to illustrate how a radically irrational belief can render certain responses rational. If M also believes that wearing a football helmet at all times (like B.D. in Doonesbury) will protect his glass head from shattering, it is rational for him to wear a football helmet at all times. There is a sense of 'rational' in which the rational thing for M to do, *given that he has those beliefs*, is to wear a football helmet at all times. And in that same sense, given the assumptions Plantinga is making, it is rational for one who holds N&E to give up R. But this sense is too subjective to have any important epistemic significance. This can be seen if we consider this noncontextually delimited question (i.e., not limited by the context of what M in fact believes): Is it rational for M to believe that whenever he takes off his helmet his head is in danger of shattering? Of course not. It is a paradigm of irrationality, and just because of the extreme irrationality of the belief from which it was derived. Similarly, if N&E is no better supported for S than R,

[14] The attentive reader of *WPF*, chapter 12, will recognize that this argument comes close to being a combination of Plantinga's "preliminary argument" (228–229) and "main argument" (beginning on 229), combining the strong points of each and avoiding their weaknesses.

then there is no rational basis for a preference for withholding R, rather than withholding N&E.

I should note that Plantinga says things that might be taken to support his raising the question "What defeats what?" in a context defined by the acceptance of N&E. In particular, he says, "In the typical case of defeat, I will first believe *b* and then later come to believe the defeater *d*" (*WCB*, 361). If he supposes this supports his procedure, he would be assuming that the temporal order of acquisition of beliefs makes an important difference to the rationality of one or another response to their incompatibility. But why suppose that? If I assumed anything along this line, which I don't, it would be that the older belief has the right of prior possession and, by a principle of conservatism, that if everything else is equal preference goes to the already established. In any event, I can't see that the fact that D is the new boy on the block gives it the right to displace B rather than vice versa, if there is no reason in comparative epistemic status to chose one response rather than another.

Finally, at the risk of undue repetition, we have already seen that Plantinga does allow comparative epistemic status (in the Aberdeen case and elsewhere) to affect the choice as to what to give up, at least where there is a significant difference in epistemic status. And reading "relevant respects" in D* in the light of those cases, why then should this (approximately) equal epistemic status not also have a bearing—in this case leaving us without rational grounds for preference, and hence leaving it ambiguous which, if either, "defeats" the other. And if it is necessary to make a choice nevertheless, why adopt the principle of always holding the newcomer fixed? Flipping a coin would have more to commend itself than that!

I think I have complained enough about Plantinga's overly subjective notion of rationality, overly subjective because too tied to what S believes, whatever the objective epistemic status of those beliefs. Now I want to point out some severe limitations of his argument vis-à-vis the conclusion he is aiming at, even if we work with his "given that S believes that D" notion of rationality. The conclusion aimed at is that an advocate of N&E cannot rationally accept R, since the former is a defeater for the latter. W can see some roadblocks in the path to this conclusion by reminding ourselves of some of the things that Plantinga recognizes as necessary for the conclusion.

1. First, the conclusion will follow only if the defeater is itself undefeated. (Or on my preferred reading, the alleged defeater will be a defeater only if undefeated.) Of course, if N&E is defeated, that will be no consolation to its advocates; and Plantinga can use this as one horn of the dilemmatic argument against N&E I suggested for him earlier. But it will affect the conclusion that one who holds N&E thereby has an (undefeated) defeater for R. And if, as Plantinga holds and as I hold, N&E is woefully undersupported, this is a live possibility, given Plantinga's views. For Plantinga holds that "with respect to some sorts of beliefs, what gives me a defeater for a belief of that sort is just

the fact that I don't have a reason for it; that realization is itself a defeater for the belief" (ND, 27).[15] I take it 'have a reason' here is to be understood as 'have a sufficient, adequate, strong . . . reason'. Moreover, N&E is certainly a thesis for which one needs sufficient *reasons* if it is to be rationally held. It could hardly be considered to be self-evident or anything like that. And, if Plantinga and I are correct, N&E *is* defeated in this way (or could be if we add Plantinga's requirement that S realizes the lack). And so this is one way in which N&E could fail to defeat R, on Plantinga's own showing.

2. Plantinga opines that even if B has a low probability on D, this is not sufficient for defeat unless S sees this D-B relationship (ND, 36–37). By imposing this requirement Plantinga makes the (alleged) probability relationship useless for reaching his intended conclusion. For presumably very few N&Eers will think that P(R/N&E) is low. Indeed, Plantinga's argument is set up in opposition to some very prominent N&Eers who affirm and argue for the opposite thesis. So with the above requirement N&E will defeat R only for those N&Eers (presumably a distinct minority) who agree with Plantinga's probability assessments. Note that if Plantinga were working with a more objective conception of rationality and hence did not make so much hang on what beliefs S actually has, he could take the objective probability relation to yield the defeat relation (if it does so) without the need to require that S believe that the probability relation holds.

3. Then there is the question of whether R has considerable support within typical noetic structures. Plantinga argues that any argument for R, indeed any argument for anything, will presuppose R, and therefore will be question begging (in a way I term 'epistemic circularity') (*WPF*, 233–234). But there are two complications here. First, I have argued elsewhere that one can have strong reasons for a reliability of faculties claim even if those reasons are the output of faculties for whose reliability they are claimed to be reasons.[16] But I won't pursue my reasons for this obviously counterintuitive claim here. Instead I will defend the possibility that R enjoys "basic warrant", to use Plantinga's term of choice. It may be that the general reliability of our cognitive faculties is a fundamental presupposition that we are all warranted in accepting in the absence of any reasons or evidence whatever. The claim that R has this status is the core of one of the objections Plantinga considers in ND, the one labeled "R Beyond Defeat". His response to that objection is not that R doesn't have warrant in this way,[17] but that it doesn't follow from R's having basic warrant that I cannot acquire a defeater for it. I agree with

[15] Note that, in accordance with his usual practice in this discussion of placing a lot of weight on what the subject in fact believes, Plantinga specifies the defeater as the *realization* that one lacks an adequate reason. I would prefer to be more objective and say that the lack of an adequate reason, whether the subject realizes it or not, could be a defeater. But let that pass.

[16] See Alston, "Epistemic Circularity," in *Epistemic Justification* (Ithaca, N.Y.: Cornell University Press, 1989).

[17] "Let's also agree that R does have warrant and perhaps a great deal of warrant, when it is taken as basic" (ND, 51).

Plantinga about that. But the crucial point is that if R does have "a great deal of basic warrant", then one would have to show that there is a defeater for it in spite of that, in order to clear the way for the conclusion that N&E defeats R. And Plantinga has not done that.[18]

IX

The only item remaining on the agenda is the other part of Plantinga's low probability–inscrutability alternative. Suppose that we are unable to determine the value of P(R/N&E). Is that sufficient, along with other relevant features of the situation, to make N&E a defeater of R?

Before tackling this question, I need to precise the inscrutability alternative. If the claim is only that we are unable to arrive at a well-grounded precise value, or even an informative range of values, that is too obvious to mention. Presumably what is intended is that we are much more at a loss than that, that we are unable to decide even between very low, somewhere in the middle half, or very high.

So what about the alternative, thus construed? Since this chapter is already so long, since the bearing of the inscrutability alternative raises many difficult questions, and since I do not at the moment see my way clear through all of them, I will limit myself to two points.

First, on the inscrutability assumption we don't have even the prima facie case for defeat that we have on the low probability assumption. If I don't know what to say about how likely or unlikely N&E makes R, how does that engender even a prima facie tension between a belief in N&E and a belief in R? Suppose I believe both a certain version of quantum mechanics (Q) and general relativity theory (T), and I have considerable evidence for each. I am inclined to think that Q has some evidential bearing on R, but I am unable to determine what that is. (This may be because one or the other is insufficiently developed, or because the links between what they apply to has not been sufficiently worked out, or because I am not capable of spelling out the connections, or . . .) In those circumstances would I have even a prima facie case that Q is a defeater for T? I can't see that I would. That would entail that unless my independent evidence for T outweighs my independent evidence for Q, then rationality requires that I give up the former. But surely that is not the case. Since I am unable to assess the probability of T on Q, I have no basis for thinking that the tenability of T depends on my having more evidence for it than for Q. What does that have to do with the rationality of accepting T?

[18] There is also the ad hominem point that Plantinga's "preliminary argument against naturalism" uses R as a premise in a Bayesian argument against naturalism (*WPF* 228–229). The fact that he feels free to do this, despite the epistemic circularity involved in any argument for R, indicates that he does not take that epistemic circularity to prevent R itself from occupying whatever favorable epistemic status is required for being properly used as a premise in arguments.

If the probability of T on Q were low, then that requirement would make sense. But when we are unable to say what that probability relation is, it would seem sheerly arbitrary to say that T has to be given up if the evidence for it is less strong than the evidence for Q.

Second, even if there were such a prima facie case for rejecting T, to make it more than prima facie it would have to be the case that T does not enjoy strong independent support, contrary to the way the case was described. But, Plantinga might well say, that is just where the N&E–R relationship differs from this case. Indeed, in giving his argument in *WPF* on the inscrutability alternative, he explicitly makes it dependent on the assumption of no independent support for R. After arguing for rationality's requiring giving up R for his "hypothetical population", he continues:

> But now suppose we again apply the same sort of reasoning to ourselves and our condition. Suppose we think N&E is true: we ourselves have evolved according to the mechanisms suggested by contemporary evolutionary theory, unguided and unorchestrated by God or anyone else. Suppose we think furthermore, that there is no way to determine P(R/(N&E&C)) (specified to us). . . . If we have no further information, then wouldn't the right attitude here, just as with respect to that hypothetical population be agnosticism, withholding belief [in R]? (*WPF*, 229)

Thus Plantinga himself makes his conclusion on the inscrutability alternative depend on the claim that there is no (sufficient) independent support for R. And, again, as pointed out in discussing the low probability assumption, he has not come close to establishing that. Hence here too the argument fails.

X

By way of summary: (1) The claim that the P(R/N&E) is low is poorly supported, as Plantinga himself admits. Taking that value to be "inscrutable" seems more reasonable, but there is no clear relevance of that to the defeat claim. (2) Even if the value of P(R/N&E) is low, it doesn't follow that N&E is a defeater of R, in any epistemically important sense of 'defeater', unless R fails to enjoy greater warrant than N&E. And it seems plausible to suppose that it does, and in any event Plantinga has failed to show that it does not.

All this leaves Plantinga's argument, as he has presented it, without the intended force. It would still be possible for him to argue that one who insists on holding on to N&E cannot preserve internal rationality except by giving up R. And perhaps that would give him the heart of what he is after. But it

doesn't give it to him in the way in which he tried to reach that result, by arguing that N&E is a defeater of R for one who holds both, nor does it yield the epistemically objective conclusion that internal rationality requires for anyone (regardless of their subjective preferences) that a conflict between N&E and R be resolved by giving up R.

Reply to Beilby's Cohorts

Alvin Plantinga

Let me repeat what I said in the initial statement of the argument: I'm delighted to express my gratitude to Jim Beilby and his cohorts for these probing and challenging comments on my evolutionary argument against naturalism (hereafter EAAN). Further, it is a privilege to be able to respond to them—a privilege that induces, if not fear and trembling, at least a healthy dose of anxiety and apprehension.[1] I should say up front, however, that

I am deeply indebted to Michael Bergmann, Tom Crisp, Trenton Merricks, Omar Mirza, Marie Pannier, Michael Rea, Ray van Arragon, and David VanderLaan for penetrating comments and wise counsel.

[1] There have been other published and semipublished objections: William Alston's comments on the paper, presented at a conference at Santa Clara University in the spring of 1992; Carl Ginet, "Comments on Plantinga's Two-Volume Work on Warrant," *Philosophy and Phenomenological Research* 55, 2 (June 1995): 403–408; Timothy O'Connor, "An Evolutionary Argument against Naturalism?" *Canadian Journal of Philosophy* 24, 4 (December 1994): 527–539; Richard Otte's comments on the paper, presented at the same symposium as Alston's; Glenn Ross, "Undefeated Naturalism," *Philosophical Studies* 87, 2 (August 1997): 159–184; David Hunt, "Is Metaphysical Naturalism Self-Defeating?" presented at the Pacific Division meetings of the American Philosophical Association, March 1994; Leopold Stubenberg, "Naturalism Undefeated," presented at a colloquium at the University of Notre Dame, 1994; Wesley Robbins, "Is Naturalism Irrational?" *Faith and Philosophy* 11, 2 (April 1994): 255–259; Evan Fales, "Plantinga's Case against Naturalistic Epistemology," *Philosophy of Science* 63, 3 (September 1996): 432–451; William Ramsey, "Plantinga Defeated?" presented at a colloquium at the University of Notre Dame, 1995; Keith Lehrer, "Proper Function vs. Systematic Coherence," in *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, ed. Jonathan Kvanvig, 25–45 (Lanham, Md.: Rowman and Littlefield, 1996); M. L. Nathan, "Naturalism and Self Defeat: Plantinga's Version," *Religious Studies* 33, 2 (June 1997): 135–142; Michael Levin, "Plantinga on Functions and the Theory of Evolution," *Australasian Journal of Philosophy* 75, 1 (March 1997): 83–98; Jerry Fodor, "Is Science Biologically Possible?" (the 1998 Benjamin Lecture at the University of Missouri); James Beilby, "Is Evolutionary Naturalism Self-Defeating?" *International Journal for the Philosophy of Religion* 42, 2 (October 1997): 69–78; Thomas McHugh Reed, "Evolutionary Skepticism," *International Journal for Philosophy of Religion* 42, 2 (October 1998): 79–96;

EAAN seems to me to emerge unscathed—or if a bit scathed, then at least bloody but unbowed. Indeed, I think the comments enable a strengthened version of the argument; some of the points made by the commentators suggest more subtle and nuanced ways of putting it. In particular, some additions or corrections have to be made to the underlying epistemology,[2] and further distinctions must be made among kinds of defeaters.

Beilby's cohorts present a large and bewildering variety of criticisms and comments. Of course I can't deal adequately with nearly all of them (I do hope you won't think I don't deal adequately with any). But in order to introduce a bit of structure, I'll first develop a couple of particularly important points, ones that come up in several of the essays; in the second part of the chapter, I'll turn to replies to individual authors.

I. General Comments

A. Humean Rationality Defeaters

Let's use the term 'naturalist' to denote someone who accepts N&E (the conjunction of metaphysical naturalism with the claim that we and our cognitive faculties have come to be by way of the mechanisms specified in contemporary evolutionary theory). Say that the *Probability Thesis* is the thesis that (R/N&E) is low or inscrutable. A central strand of EAAN is the claim that the naturalist who sees the truth of the Probability Thesis has a defeater for R. The defeater in question, according to me, is a *rationality* defeater: the reflective naturalist—one who recognizes the Probability Thesis—will be in some sense irrational if he continues to believe R. In *what* sense? At the fundamental level, I explain rationality in terms of proper function: the rational thing to believe, in circumstances C, is what a properly functioning human being—more exactly, one whose cognitive or rational faculties are functioning properly in the relevant respects—would believe in those circumstances. But now suppose I'm a naturalist and come to see that the Probability Thesis is in fact

---

Branden Fitelson and Elliott Sober, "Plantinga's Probability Arguments against Evolutionary Naturalism," *Pacific Philosophical* Quarterly 79 (1998): 115–129; Anthony Peressini, "Naturalism, Evolution, and Self-Defeat," *International Journal for Philosophy of Religion* 44, 1 (August 1998): 41–51; and David Silver, "Evolutionary Naturalism and the Reliability of Our Cognitive Faculties," *Faith and Philosophy*, forthcoming. There is also a response by Eric Kraemer at the Central Division meeting of the American Philosophical Association, April 1996; a paper by Elliott Sober at the Chapel Hill Philosophy Colloquium, October 1997; and papers by Bill Ramsey, Ric Otte, and Michael Bergmann at a World Congress of Philosophy symposium, August 1998. The papers by Fodor, Ramsey, Otte, and Bergmann are ancestors of their chapters in the present volume. In the course of replying to the chapters in this volume, I shall also respond to some of the objections raised in the other papers mentioned above.

[2] This is, fundamentally, the epistemology of *Warrant and Proper Function* (New York: Oxford University Press, 1993) (hereafter *WPF*); EAAN, however, does not presuppose this or any other specific epistemology.

true. Doesn't it seem that I would nonetheless continue to believe R, that my cognitive faculties are reliable? Won't I at any rate continue to *assume* this, in the sense that when I am not explicitly thinking about this question, I will carry on my intellectual life as I did before, accepting the testimony of my senses and memory, reasoning in accord with *modus ponens*, and the like? Indeed, wouldn't the same go even if, on an injudicious reading of Descartes, I come to believe that I have been created by an evil demon who takes special delight in creating creatures who think they are the epistemic lords of the universe but in fact hold absurdly false beliefs; or even if (to turn to the contemporary version of Descartes's fantasy) I come to believe that I am a brain in a vat, serving as a subject of cognitive experiments carried out by evil scientists from a highly advanced alien civilization? Won't I carry on in the same way even if I become convinced I've contracted mad cow disease, a disease, as I believe, that renders its victims cognitively unreliable? And (the crucial point) in so doing, might I not be functioning perfectly properly, without so much as a hint of dysfunction or malfunction?

Clearly the answer to both questions is yes; I would carry on as before, and that, presumably, is exactly what is required by proper function. But then doesn't it follow that I am mistaken in claiming that the naturalist who accepts the Probability Thesis has a rationality defeater for R? It looks as if you don't get a rationality defeater for R even if you believe that you are a victim of a Cartesian evil demon, or of Alpha-Centaurian cognitive superscientists, or of mad cow disease. A fortiori, then, you won't get such a defeater just by accepting *N&E & P(R/N&E) is low or inscrutable*. So what do I have to say for myself?

This objection, first called to my attention by William Talbott, is both formidable and revealing—revealing, in that it lays bare important further considerations about defeaters. Let's turn to an example of the sort Talbott proposes. Suppose you believe that XX is a drug that causes widespread cognitive disorder: an hour after XX is ingested, the subject's cognitive function is no longer reliable. Add that the drug also prevents its victim from detecting her unreliability. Now suppose someone believes that she has taken this drug, and suppose she also believes that 90 percent of those who take the drug are no longer reliable, although of course unable to detect their unreliability. The first thing to see here is that a person in this condition will continue to form perceptual beliefs, memory beliefs, and the like; and she will also continue, in some sense, to assume that her cognitive faculties are reliable. Furthermore (at least so one thinks), in so doing she is conforming to our design plan; this is precisely what proper function requires. In this sense, then, her cognitive behavior is entirely rational.

Still, one wants to say that in *some* way she really *does* have a defeater for R specified to her case. She believes that R no longer holds for the vast majority of those who take this drug; she knows of nothing that distinguishes her case from theirs; someone who was thinking straight would certainly no longer

(without additional evidence) believe R for someone *else* she thought had ingested XX; so doesn't rationality dictate that she cease believing R in her own case?

Here we must go back to one of the necessary conditions of warrant. In *Warrant and Proper Function* I argued that a belief's being formed by cognitive faculties functioning properly in the appropriate epistemic environment is necessary but not sufficient for warrant (*WPF*, 11–17). That is because our cognitive establishment contains (or at any rate could contain) processes, modules, whose purpose it is not to produce true belief, but belief with some other property. If you fall victim to a usually fatal disease, you may somehow think your chances are much better than is indicated by the statistics you know ("the optimistic overrider"); a woman may forget completely about the pains of childbirth; perhaps most of us think more highly of ourselves than the facts we know warrant; you may continue to believe in a friend's honesty long after a cool look at the facts would have dictated a reluctant change of mind. Another example, this one from a Freudian perspective. We human beings find ourselves thrown into a cold and indifferent world, says Freud; nature cares not a whit about us; she inflicts on us distress, pain, suffering; and she ultimately demands our death. The generality of humankind can't directly face this situation (Freud evidently thought himself an exception); life without illusion would cause depression, apathy, debility. We human beings therefore (unconsciously, of course) project into the heavens a Father Who has created us and the world, Who is entirely in charge of whatever happens, and Who loves us, His children. Our forming this belief in this way is not, according to Freud, to fall victim to cognitive malfunction; indeed, this is precisely what proper function requires. The bit of functioning in question, however, is not truth-aimed; it is not aimed at the production of *true* belief, but rather belief enabling us to carry on in this cruel and heartless world.

These beliefs, therefore, do not issue from cognitive processes whose purpose it is to produce true beliefs; their purpose is instead to produce beliefs with some other useful property: willingness to have more children, or ability to form serious friendships, or likelihood of recovery, or (says Freud) ability to carry on in this cold and unfeeling world. But then being produced by properly functioning cognitive capacities is not sufficient for warrant; in addition, to have warrant, a belief must be such that it is produced by a module that is successfully aimed at truth. The above beliefs do not meet this condition, although they do meet the proper-function and congenial-environment condition. They display proper-function rationality (they are produced by properly functioning rational or cognitive faculties) but they do not have warrant.

This distinction between proper-function rationality and warrant (or a near neighbor) must also be made with respect to defeaters. On some occasion a person might acquire powerful evidence against a given belief *B* but fail to change belief, due, not to dysfunction, but to the fact that the processes

producing *B* are aimed not at the production of true belief (that is, not at the maximization of true belief and the minimization of false belief) but at the production of belief with some other property. In such a case, the design plan may dictate no change of belief in the face of the current input—not in the service of an aim to maximize true belief and minimize false belief, but in order to accomplish some other end. So consider someone who believes he will live for many years, but then gets the bad news that he has an ailment for which the statistics indicate he probably has only two months to live. The optimistic overrider may dictate that he continue to believe he'll live for much more than a couple of months, and this may be in accord with his design plan. Still, I'd say he does have a defeater of *some* sort for the belief that he will live for a long time, because the belief he actually forms is incompatible with beliefs that would result from processes aimed at the production of *true* belief.

The case of ingesting XX is of this very kind. Given our cognitive architecture, we can't properly carry on our epistemic lives unless we believe or assume that our cognitive faculties are reliable. More exactly, given our cognitive architecture, we can't carry on our epistemic lives if we explicitly *disbelieve* or *seriously doubt* that our faculties are reliable. So any reason to doubt or to disbelieve R will be overridden by the imperative force of this requirement. But then the overriding force does not come by way of a direct aim at the formation of *true* belief; and hence in *some* sense the person who believes she has ingested XX really does have a defeater for R, even though proper function dictates that she continue to assume that R is true. *What* sense? The clue is as follows: if only our truth-aimed processes were at work in these situations, she *would* have a defeater. Some of the processes governing the maintenance of R are aimed not directly at truth but rather at the ability to carry on our cognitive life. And our design plan is such that if those processes—the processes that govern the maintenance of R but are not directly aimed at truth—were excised, she *would* have an ordinary rationality defeater. This becomes evident when we reflect on what we would think about the reliability of *someone else* who had ingested XX.

In "Naturalism Defeated"[3] I defined defeat (rationality defeat) as follows:

> (D) *D* is a defeater of *B* for *S* at *t* iff S comes to believe *D* at *t* and *S*'s noetic structure *N* at *t* includes *B* and is such that any human being (1) whose cognitive faculties are functioning properly in the relevant respects,[4] (2) whose noetic structure is *N*, and (3) who comes to believe *D* but nothing else independent of or stronger than *D* would withhold *B* (or believe it less strongly).[5]

[3] Unpublished but available on the web at www.homestead.com/philofreligion/files/alspaper.htm.

[4] For comment on what respects are relevant here, see my reply to Alston below.

[5] We could use the term 'partial defeater' for defeaters that require believing *B* less firmly. A full treatment would explain degrees of belief (which are not to be thought of as probability judgments; see *WPF*, 166) and show how partial and full defeat are related. Here there is no space for that: but note that defeat is really a special case of partial defeat, at least if we stipulate that coming to withhold

(D) is a perfectly sensible account of one kind of rationality defeater: *proper-function* rationality defeaters (proper function indexed to the human design plan). But there is also another kind that is important in the present context: one we could call a *purely alethic* rationality defeater. Say that a cognitive process aimed directly at the production of true beliefs is an alethic process. As I've just been arguing, there are both alethic and non-alethic processes. Now some defeaters are such that only alethic processes are involved in their operation, as when you've always thought there weren't any cacti in the Upper Peninsula of Michigan, but then see a luxuriant stand of prickly pear cactus on a trip to Iron Mountain. That would be a purely alethic defeater; but how can we characterize this notion of purely alethic defeat? Note first that a defeater is a defeater *with respect to a certain way of working, a cognitive design plan*. We've implicitly been confining the discussion to the *human* design plan, but there are many other cognitive design plans, some instantiated and some not; and what is a defeater with respect to one of them may not be with respect to another. Now consider a cognitive design plan *C* that is just like the human design plan with respect to all purely alethic processes, but contains no non-alethic processes. A purely alethic defeater (with respect to the human design plan) for a belief *B* is a defeater for *B* with respect to *C*. We could also put it as follows (this definition is from *Warranted Christian Belief*, where I used 'purely epistemic' for 'purely alethic'):

> (D) *D* is a purely epistemic defeater of *B* for *S* at *t* iff (1) *S*'s noetic structure *N* at *t* includes *B* and *S* comes to believe *D* at *t*, and (2) any person *S** (a) whose cognitive faculties are functioning properly in the relevant respects, (b) who is such that the bit of the design plan governing the sustaining of *B* in her noetic structure is successfully aimed at truth (i.e., at the maximization of true belief and minimization of false belief) and nothing more, (c) whose noetic structure is *N* and includes *B*, and (d) who comes to believe *D* but nothing else independent of or stronger than *D*, would withhold *B* (or believe it less strongly).[6]

Thus according to Freud, theistic belief is subject to a purely alethic rationality defeater. The person who believes she has ingested XX has a rationality defeater of this kind for R; she has a purely alethic rationality defeater for that belief. More important in the present context, my claim is that the naturalist who comes to see that P(R/N&E) is low or inscrutable also has a purely alethic rationality defeater. Granted: he will continue to assume R; nevertheless he has a defeater of this sort.

But he also has a certain sort of ordinary proper-function rationality defeater. For consider what happens when I reflect on my cognitive situation after having ingested XX. I am aware of the fact that I continue to assume R

B is a special case of coming to believe *B* less strongly. For the sake of brevity I will henceforth suppress mention of partial defeaters, although the application of what I say to them should be routine.

[6] *Warranted Christian Belief* (New York: Oxford University Press), 363.

in the face of my belief that I have ingested XX—what else can I do?—and this in the face of evidence that would lead me to withhold R with respect to anyone else. I also realize that I don't know of anything that distinguishes my case from theirs; and I come to see that what underlies my continuing to believe R in this way is just the fact that the relevant bit of the design plan is aimed not at truth but at the avoidance of cognitive disaster. Then I will no longer believe R, or at any rate will not believe it in an unqualified way. My situation will be like that of Hume, who concludes that while, to be sure, he can't stop believing what nature prompts him to believe, these beliefs are at best doubtful. When he follows out what seem to be the promptings and leading of reason, he winds up time after time in a black coalpit, not knowing which way to turn:

> Where am I, or what? From what causes do I derive my existence, and to what condition shall I return? Whose favour shall I court, and whose anger must I dread? What beings surround me? and on whom have I any influence, or who have any influence on me? I am confounded with all these questions, and begin to fancy myself in the most deplorable condition imaginable, inviron'd with the deepest darkness, and utterly depriv'd of the use of every member and faculty.[7]

Of course this is Hume in his study, sometime before he emerges for that famous game of backgammon. Nature herself, fortunately, dispels these clouds of despair: she "cures me of this philosophical melancholy and delirium, either by relaxing this bent of mind, or by some avocation, and lively impression of my senses, which obliterate all these chimeras. I dine, I play a game of back-gammon, I converse, and am merry with my friends." Still, the enlightened person, Hume thinks, holds the consolations of Nature at arm's length. She knows she can't help acquiescing in the common illusion, but she maintains her skepticism of "the general maxims of the world" and adopts a certain ironic distance, a wary double-mindedness: "I may, nay I must yield to the current of nature, in submitting to my senses and understanding; and in this blind submission I shew most perfectly my sceptical disposition and principles."[8] This is the irony of the human condition: those who are enlightened can see that what nature inevitably leads us to believe is false, or arbitrary, or at best extremely dubious; but they also see that even the best of us simply don't have it in them to resist her blandishments. We can't help believing those "general maxims," or if we can, it is only for brief periods of time and in artificial situations. No one can think Humean thoughts about induction, say, when under attack by a grizzly, or when unroped and clinging desperately to a rock face high above the valley floor. You won't find yourself saying, "Well, naturally I can't help believing that if this handhold breaks out I'll hurtle down to the ground and smash into those rocks, but (fleeting, sardonic, self-

[7] *Treatise*, ed. Selby-Bigge (Oxford: Clarendon Press, 1998), 269.
[8] Ibid.

deprecatory smile) I also know that this thought is just a deliverance of my nature and is therefore not really to be taken seriously." Still, in other circumstances one can take a sort of condescending and dismissive stance with respect to these promptings of nature; in reflective moments in my study, for example, I see through them. As a rational creature I can rise above them, recognizing that they have little or nothing to be said for them.

That's the situation the naturalist will be in: he won't be able to help believing or at least assuming R; but (if he reflects on the matter) he will also think, sadly enough, that what he can't help believing is unlikely to be true. He will have a purely alethic rationality defeater for R, but at those reflective moments when he thinks about his cognitive situation he will also have a proper-function rationality defeater for R—let's call it a *Humean* rationality defeater. Of course when he *isn't* thinking about his epistemic condition he will naturally assume that R is true—nature permits nothing else. But when he reflectively considers his condition he will, if functioning properly, adopt that Humean posture, or something similar. He has a Humean defeater. This introduces a qualification into my argument, a qualification that strengthens it by adding appropriate nuance. And this also requires a modification of the underlying epistemology: it requires the recognition of purely alethic rationality defeaters and also Humean defeaters.

B. Semantic Epiphenomenalism

Jerry Fodor expresses considerable anguish at the thought that perhaps the *content* of belief does not figure in the causal chain leading to behavior:

> I'm not really convinced that it matters very much whether the mental is physical; still less that it matters very much whether we can prove that it is. Whereas, if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying . . . if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.[9]

Right: this (call it 'semantic epiphenomenalism') is indeed a haunting specter for a materialist, so haunting as to tempt him to ignore this possibility altogether (as Fodor does in his contribution to this volume). But of course that hardly shows that the probability of semantic epiphenomenalism on N&E is low. Semantic epiphenomenalism fits into EAAN as follows. A crucial premise of EAAN is that P(R/N&E) is low or inscrutable. Suppose we argue this by breaking N&E into three mutually exclusive and jointly exhaus-

[9] Jerry Fodor, "Making Mind Matter More," in *A Theory of Content and Other Essays* (Cambridge: MIT Press, 1990), 156.

tive subcases;[10] P(R/N&E) will be the weighted average of P(R/N&E&S_i) for each of the three S_i (weighted by the probability of the S_i in question on N&E). One of the three possibilities is the hypothesis that belief does not in any way enter into the causal chain that leads to behavior; another is that belief *does* enter into the causal chain leading to behavior, but only by way of its 'syntactic' (physiochemical, neurophysiological) properties—not by way of its semantic properties such as truth or falsehood, intentionality, and, in particular, content; and the third possibility is that belief is causally efficacious with respect to behavior, both by way of neurophysiological properties and by way of content. Now it is fairly clear, I think, that P(R/N&E & semantic epiphenomenalism) is low (or possibly inscrutable). If P(semantic epiphenomenalism/N&E) is high, then P(R/N&E) will be low (so far). In what follows I shall explain semantic epiphenomenalism more fully and argue in more detail that its probability on N&E is high or inscrutable. I can't treat this subject at the length it deserves—it surely deserves a book all of its own—but I think we can enter the subject far enough to see that the probability of semantic epiphenomenalism on N&E is indeed high (or inscrutable).[11]

Well, why think P(semantic epiphenomenalism/N&E) *is* high (or inscrutable)? Presumably one who accepts metaphysical naturalism will be a materialist with respect to human beings; at any rate all the contemporary naturalists I know are also materialists. So let's at least temporarily assimilate materialism with respect to human beings to naturalism. Now suppose materialism is true: given that there *are* such things as beliefs, what sort of thing will a belief *be*?[12] Since we are assuming materialism, they will presumably have to be material processes or events—perhaps long-standing neural events of some kind in the nervous system.[13] Perhaps a belief will be a neural event involving some of the approximately 10^{11} neurons most of us enjoy; perhaps it will involve these neurons firing in a certain sort of pattern, with inputs and outputs to and from other neural processes and events. So considered, beliefs will of course be able to enter the causal process that leads to behavior.

So considered, however, beliefs will have two quite different sorts of prop-

[10] As opposed to the four subcases in the argument as presented in the initial statement of the argument. (Here I am ignoring the third of those possibilities, the one according to which belief causes behavior, by virtue of semantic properties as well as neurophysiological properties, but is maladaptive.)

[11] The argument I'll propose is related in ways that aren't entirely clear to arguments by Jaegwon Kim in *Mind in a Physical World* (Cambridge: MIT Press, 1999) and "The Myth of Non-reductive Materialism," in *Supervenience and Mind* (Cambridge: Cambridge University Press, 1993). I am also indebted here to Trenton Merricks.

[12] A materialist might, of course, hold that human beings sometimes display the property of believing *p*, for some proposition *p*, while denying that there are any such things as beliefs. For what follows, this difference will make no difference.

[13] Note that if property dualism is true, these events could be mental events in that they (or the objects involved in them) display mental properties; they are still material events, however, in that, of course, only material substances will be involved in them.

erties. As we saw above, such an event will have neurophysiological properties. But if this neural event is really a *belief*—perhaps the belief that Proust was a more subtle writer than Mickey Spillane—it must have another kind of property as well. It will have to be the belief that *p* for some proposition *p*—in this case the proposition that Proust was more subtle than Spillane. Some proposition *p* will be uniquely associated with it as its *content*: the proposition *p* such that the belief in question is the belief that *p*. And by virtue of having content, it will also have intentionality or aboutness; the proposition that is the content of the belief will be directed on some subject matter (Proust and Spillane, say), and predicate something of that subject matter (that the former is more subtle than the latter); the belief shares this intentionality. So if materialism is true, then beliefs, if there are any such things, will be neural structures that display these two kinds of properties.

Now as we all know, it is far from clear how a neural structure can acquire a content. How does a neural event somehow get assigned a certain proposition as its content? It is hard to think of any scenarios that are as much as decently plausible.[14] But (again, if materialism is to accommodate the idea that there really are such things as beliefs) it must happen somehow. Granted that such a structure can acquire a content, however, it still seems extremely difficult to see how it can enter the causal chain leading to behavior *by virtue of its content*. It is easy enough to see how it can enter the causal chain by virtue of its neurophysiological properties. We learn from the neuroscientists that our limbs move in response to a volley of electrical impulses originating in the central nervous system. Emanating from the brain or elsewhere, these electrical impulses are propagated along certain channels, the *efferent* nerves; when they reach the muscles, they cause contraction; and of course it is only by virtue of muscle contraction that motor output takes place. So it is clear how a belief, a neural state or structure or event, can play a role in the causation of behavior by virtue of its neurophysiological properties. But how does the *content* of this belief, that proposition, enter that causal chain? As Fred Dretske puts it,

> We can, following Davidson, say that reasons *are* causes, but the problem is to understand how their being reasons contributes to, or helps explain, their effects on motor output. The fact that they have a content, the fact that they have a *semantic* character, must be relevant to the kind of effects they produce. If brain structures possessing meaning affect motor output in the way the soprano's acoustic productions affect glass, then the meaning of these neural structures is causally inert. Even if it is there, it doesn't *do* anything. If having a mind is having *this* kind of meaning in the head, one may as well not have a mind.[15]

[14] In *Explaining Behavior* (Cambridge: MIT Press, 1988) Fred Dretske makes a valiant but, I think, wholly unsuccessful attempt to provide such a scenario.

[15] Ibid., 80.

Dretske goes on to note that this problem is clearly an analogue of the one often thought to afflict Cartesian dualism: the problem of explaining how immaterial events or structures (*beliefs*, for example, which on a dualistic view will be modes of an immaterial substance) can causally affect the hard, heavy, massive material world. For the materialist, the analogous problem is that of specifying how a neural structure's propositional *content* can causally influence behavior. When the soprano hits the high C and shatters the champagne glass, it is not by virtue of the content or meaning of the line she sings that the glass is shattered; it is just by virtue of the physical properties of the event in question. Even if the words she sings had a wholly different and indeed contrary content—instead of meaning "Oh how I love him!" they mean "Oh how I hate him!"—the glass would shatter in the same way. The content seems to be causally impotent, causally irrelevant. And how can the materialist avoid the same conclusion in the case of the content enjoyed by those neural structures that are beliefs? Suppose a given neural structure both plays a causal role in the production of motor input and also has a certain content: that Proust is more subtle than Spillane. It seems that the causal role it plays depends just on the neurophysiological properties it possesses: would it not produce the same physical effect, the same effect on nerves and muscles and glands, even if it had a quite different content, indeed, even if its content were the proposition that Spillane is more subtle than Proust?[16] From the perspective of materialism, therefore, semantic epiphenomenalism is hard to avoid. This means initially that our estimate of the probability of semantic epiphenomenalism on N&E or at any rate on N&E&M (materialism with respect to human beings) should be high.

But it also seems that P(R/N&E & semantic epiphenomenalism) is low (or perhaps inscrutable). We can see this as follows. Given a certain array of behavior-producing structures, natural selection can modify these structures in the direction of greater fitness or adaptiveness, according a reproductive edge to those organisms sporting mutations (mutations involving these structures) that permit or enhance reproductive success. The same will go for those neural structures that produce beliefs. Natural selection can eliminate certain structures (those that come at the cost of reproductive fitness) and encourage others (those that enhance reproductive fitness). If false belief caused maladaptive action, therefore, natural selection could presumably modify belief-producing structures in the direction of greater reliability—a greater proportion of true as opposed to false belief. But if content does not enter the causal chain that leads to behavior, then of course it will not be the case that a belief causes maladaptive behavior by virtue of its being false, and it will not be the case that a true belief causes the behavior it does by virtue of being true. And

[16] Note that if content *supervenes* on neurophysiological properties (together, perhaps, with certain environmental properties) then it won't be possible that this event have the same neurophysiological properties but a different content.

then it is hard to see how natural selection can promote or enhance or reward true belief and penalize false belief.

Now the *reliability* of a cognitive faculty or process requires the truth of at least a substantial proportion of the beliefs it produces.[17] But suppose semantic epiphenomenalism is true. Then by modifying behavior (more exactly, the structures that produce behavior) in the direction of greater fitness, natural selection would not be modifying belief (the structures that produce belief) in the direction of greater reliability, that is, greater proportion of true belief. How then should we think of the probabilities involved? We could proceed in more than one way. First, we might ask about the probability of a *given* belief's being true, on N&E & semantic epiphenomenalism. Well, about all we can claim to know here is that this belief is probably fitness enhancing by virtue of its neurophysiological properties, and has been produced by structures that are fitness enhancing. But this tells us nothing at all about the likelihood of its *truth*. Presumably, therefore, we should think that this belief is about as likely to be true as to be false; this probability would seem to be in the neighborhood of .5. But if so, then it will be monumentally improbable that the whole structure of beliefs displays the preponderance of true belief over false that reliability requires. If I have a thousand beliefs, for example, the probability (under these conditions) that three-quarters or more of these beliefs are true (certainly not an excessive requirement for reliability) will be less than 10^{-58}.[18] And even if I am running a modest epistemic establishment of only a hundred beliefs, the probability that three-quarters of them are true, on chance alone, is something like .000001. Looking at the matter from this perspective, we should judge that P(R/N&E&SE) (where SE is Semantic Epiphenomenalism) is very low indeed.

On the other hand, we might think that *one* relevant possibility is that content gets attached to belief in this random fashion, but another is that it is somehow mainly *true* content, true propositions, that get attached to beliefs with adaptive neurophysiological properties, and still another is that it is mainly false content that gets so attached. If you think these possibilities are about equiprobable, then the thing to think is that P(R/N&E&SE) is in the neighborhood of .33. But you might also think that we have no sensible way at all of estimating these probabilities, in which case you should take P(R/N&E&SE) to be inscrutable.

So it looks initially as if P(SE/N&E) is high and P(R/N&E&SE) is low or inscrutable. This is how it looks *initially;* appearances are sometimes misleading, however: is there some way in which N&E&M can be reconciled with the idea that the content of belief plays a role in the causation of behavior? Is

[17] What reliability really requires is that the ratio of true beliefs to false, integrated over the appropriate nearby possible worlds, be high. Accordingly, it is *possible* that our cognitive faculties are reliable but produce mostly false beliefs in alpha, the actual world—possible, but very unlikely. •

[18] My thanks to Paul Zwier, who performed the calculation.

there any way in which the materialist can ward off semantic epiphenomenalism? Perhaps. We can look into this possibility by starting with another of Jerry Fodor's anxieties:

> The worry about representation is above all that the semantics (and/or the intentional) will prove permanently recalcitrant to integration in the natural order; for example that the semantic/intentional properties of things will fail to supervene upon their physical properties. What is required to relieve the worry is therefore, at a minimum, the framing of *naturalistic* conditions for representation. That is, what we want at a minimum is something of the form '*R represents S*' *is true iff C* where the vocabulary in which condition C is couched contains neither intentional nor semantical expressions.[19]

Here the hope is that *semantics* will be *determined by* syntax: that the content of a belief will be determined by its neurophysiological properties, or perhaps by the neurophysiological properties of the belief together with the properties of other relevant structures, or perhaps by the physical properties of the whole organism together with relevant properties of the environment (wide content); in this way the intentional would be integrated into the natural order.

This could happen in at least two ways. First, it could be that a certain constellation of neurophysiological and other physical properties *just is* a certain semantic property. A belief will have as its content the proposition that Proust is more subtle than Spillane just if that belief (that neural structure) displays a certain constellation of neurophysiological properties: because the property of having that content *is identical with* the property of displaying that constellation of neurophysiological properties. This could be so, if, for example, the property of having that content were identical with some Boolean combination of neurophysiological properties: where $P_1, \ldots, P_n$ are the relevant neurophysiological and other physical properties, perhaps the property *having as content the proposition that Proust is more subtle than Spillane* is the disjunction of a certain set of conjunctions of the P_i. This alleged possibility is a little hard to envisage—is it even conceivable that there is a combination of neurophysiological (and other physical) properties such that the property of having *that* combination of properties would *just be* the property of having as content that Proust is more subtle than Spillane?—but for present purposes we can set aside the initial implausibility of the suggestion.

Alternatively (and this seems to be what Fodor is envisaging), it could be that content properties—properties like *having as content the proposition that Proust is more subtle than Spillane*—are *emergent* properties: they aren't identical with any Boolean combination of neurophysiological properties, but instead *supervene* on properties of this sort. This supervention could be either

[19] Fodor, "Semantics Wisconsin Style," in *A Theory of Content and Other Essays* (Cambridge: MIT Press, 1990), 32.

broadly logical or nomic. It might be that upon achieving a certain level of neurophysiological complexity, a neuronal structure of the relevant kind simply acquires a content; its content simply appears. On this suggestion, there will be certain Boolean combinations of physical properties that entail (in the case of broadly logical supervention) or nomically imply (in the case of nomic) the property of having the content in question.[20] No doubt this suggestion is speculative *in excelsis*, but for present purposes we needn't let that interfere.

We have three ways, then, in which content "can be integrated into the natural order," as Fodor says; the case where the content property *is identical* with a Boolean combination of physical properties, the case where the first is *entailed* by the second, and the case where the first is *nomically implied* by the second. These address that worry that the intentional "will prove permanently recalcitrant to integration in the natural order." Still further, perhaps we could also see, here, a way in which semantic epiphenomenalism could be eluded. For in all three cases there may be generalizations, perhaps even lawlike generalizations, linking content with motor output.[21] In each of these cases, there will be a function F from combinations S of physical properties to content properties. Now suppose a given content property C—i.e., the property of having p as content, for some proposition p—is the value of the function for a given combination S of physical properties; and suppose that is the only combination S such that $F(S) = C$. Then it will be at least nomically necessary that an organism displays S if and only if it displays C. Accordingly, consider any generalization G in which S figures: C can replace S in G *salva veritate*, and indeed *salva modalitate* (if there is any such thing)—broadly logical modality, if it is identity or broadly logical supervention that is at issue, and physical or nomic modality otherwise. If there are generalizations linking S with motor output, therefore, there will also be generalizations linking C with motor output—linking the holding of beliefs with a particular content with displaying a particular motor output. So in this way the materialist might claim that if content is identical with certain physical properties, or supervenes either weakly or strongly on such properties, then content is causally relevant to behavior, or motor output, after all.

Now I do not believe that the supervenience of content properties on physical properties is sufficient for the former's entering the causal chain leading to behavior: it is still the physical properties by virtue of which the belief in question makes a causal contribution to behavior. It could be nomologically necessary that whenever a property A is instantiated, so is B, when in fact it wasn't by virtue of being A that A events cause B events. For perhaps there is a property X such that it is by virtue of X (and not A) that A events cause B

[20] I'm thinking of a property P as entailing (or nomically implying) a property Q just if a thing's having P entails (or nomically implies) its having Q.

[21] See Fodor, "Making Mind Matter More," 156.

events, but it is nomologically necessary that a thing has *A* if and only if it has *X*. What property is it by virtue of which a thrown brick causes a window to break? Whatever it is, call it *X*. It could be nomologically necessary that all and only white things have *X*: it wouldn't follow that thrown bricks cause windows to break by virtue of being white. It's *X* that counts here, not being white. Suppose there is a physical property *P* such that, necessarily, a human being has *P* just when she thinks of a horse; and suppose having *P* causes her to have *Q*, some other physical (maybe behavioral) property: it doesn't follow that her thinking of a horse causes her to have *Q*, or that it is by virtue of its content that this thought causes her to have *Q*. In these cases of supervention, it is still really the neurophysiological properties, not the content, in virtue of which the beliefs, the neural structures in question, enter into the causal chain leading to behavior.

But there's no need to argue about this; it really doesn't matter. That is because in all three of these cases, just as with respect to semantic epiphenomenalism, the relevant probability of R will still be low or inscrutable. Abbreviate 'content properties are identical with or supervene on neurophysiological and other physical properties' by 'S': then P(R/N&E&S) will be low or inscrutable. We can see this as follows. Take first the identity case: the case where the content property just is a combination of neurophysiological properties. Now consider a neural structure *S* that is a belief, and suppose *S* displays the combination *P* of physical properties that is identical with the property of having *p* as content, for some proposition *p*. By hypothesis, *S*'s being in *P* is identical with *S*'s having *p* as content. Suppose, furthermore, that *S*'s being in *P* is adaptive, fitness enhancing. It follows, of course, that having *p* as content is fitness enhancing by virtue of the fact that *P* contributes causally to fitness-enhancing behavior. Now what is the likelihood that *p*, the content in question, is in fact *true*? Well, there isn't the slightest reason to think, so far, that it is more likely than not that *p* is true. Take a neurophysical property *P* that is in fact adaptive, and is also identical with the property of having content *p*: so far as adaptivity goes, it doesn't in general matter whether *p* is true or false. *P* would have the same adaptive effects, one thinks, if *p* were false. How *P* contributes to motor output doesn't depend (except in special cases) upon the truth value of *p*.

And the same thing happens if the materialist proposes, perhaps more plausibly, that content *supervenes* (causally or logically) on neurophysiological properties, but is not reducible to them in the above sense. Here too there will be a function from combinations of neurophysiological properties to content properties. Now suppose a given combination *C* of neurophysiological properties is in fact adaptive, fitness enhancing. And suppose further that a given content *p* supervenes on *P*; that is, it is necessary, either nomically or in the broadly logical sense, that an organism that displays *P* will also believe *p*. What is the likelihood that *p* is true? Again, the most reasonable guess, it seems, is that this probability will not be far from ½. So far as truth of super-

vening belief is concerned, natural selection will simply have to take potluck. It selects for neural structures with adaptive neurophysiological properties; but then it must just put up (not that it minds one way or the other) with the belief content, whatever this happens to be, that goes along with those neurophysiological properties. As far as adaptivity goes, it doesn't matter whether the supervening content is true or false. It would be no more than a sort of touching natural piety to suppose that because the neurophysiological properties are adaptive, the supervening content properties must involve true propositions. P(R/N&E&S) will be low, or at best inscrutable.

So suppose there are lawlike correlations between content and motor output: that doesn't mean that natural selection can modify belief-producing structures in the direction of reliability. It modifies behavior-producing structures in the direction of greater adaptiveness, greater fitness enhancement; in so doing, it also, presumably, modifies content, since different combinations of neurophysical properties may entail different content; but there isn't the slightest reason to think that modifying behavior-producing structures in the direction of adaptiveness will also be modifying content in the direction of truth. On N&E&S it seems the best guess is the probability of a given belief's being true will be fairly close to .5; and the probability that, say, 75 percent of the whole ensemble of beliefs are true will be very low. It would be something of a cosmic miracle (or maybe cosmic magic, or an enormous cosmic coincidence) if adaptivity with respect to neurophysiological properties were correlated with truth of associated content. Again, of course, we may be suspicious of this way of assigning probabilities, claiming that all we can really say, here, is that the probabilities in question are inscrutable for us. But in either event the conclusion to draw is that P(R/N&E&S) is either low or inscrutable.[22]

Now at this point a certain bookkeeping problem arises. We began by considering P(R/N&E & semantic epiphenomenalism), which we took to be low or inscrutable. We then noted that someone (Fodor and Pettit, for example) might claim that perhaps content properties are identical with combinations of neurophysiological properties, or supervene on such properties, in which case we can plausibly say that content enters the causal chain leading to behavior after all. I said I didn't think supervention was sufficient, but in any case P(R/N&E&S) (S for supervention or identity) is the same as P(R/N&E & semantic epiphenomenalism). For purposes of present bookkeeping, therefore, I shall annex supervention or identity to semantic epiphenomenalism.

But then how shall we think about the remaining case, the case where content is in fact causally efficacious with respect to behavior, and does not supervene on neurophysiological properties (together, perhaps, with certain properties of the environment)? I suppose something like this is the com-

[22] This is so for *beliefs*. Beliefs are a special case of *representations;* and we might think things go differently for some cases of representations, in that if *they* supervene on neurophysiological properties, then their content would have to be in some sense accurate. I take this topic up in the reply to Fales.

monsense alternative. We ordinarily assume that our beliefs do indeed causally influence our behavior by way of their content; and on reflection we are also inclined to think that this is not just by virtue of the content's going along for the ride with respect to neurophysiological properties. Now one way in which this could happen is provided by substance dualism. Here the self or person or ego is taken to be an immaterial object that is nevertheless capable of causing changes in the body and hence in the material world. Still further, on this way of thinking, it is natural to take it that the self displays a certain independence of or freedom with respect to the body and the material world generally, so that it is not the case that such properties as believing *p* or hoping for *q* or fearing that *r* supervene on properties of the body or the body together with other parts of the material environment. But of course substance dualism, even if endorsed by common sense, is not popular with naturalists, and (naturally enough) even less popular with materialists.

So how can the materialist think of this last possibility—the possibility (call it 'C') that content is causally efficacious but does *not* supervene on physical properties? I can't see how this could go. Beliefs would be neural events of some sort; they would acquire content in some mysterious way; by virtue of that content, belief would make causal contribution to behavior; but it would not make that contribution by virtue of *supervention*. I can't see how this could go, and neither, I think, can anyone else. This possibility, I believe, requires serious dualism. I suspect the objective probability of this possibility on N&E (N still taken to include materialism about human beings) is zero. But suppose we concede that C is indeed a possibility, given materialism. (To borrow a line from Fodor, it is very nice of me to concede this, since I certainly don't believe it.) Still, the probability of C on N&E will still have to be estimated as very low (or inscrutable); that is because no one can think of a way in which it could happen. Accordingly, suppose we assimilate the supervention case to semantic epiphenomenalism and abbreviate it by '–C'; then P(–C/N&E) will be high (or inscrutable) and P(C/N&E) low or inscrutable. But P(R/N&E&–C) will be low or inscrutable. Therefore P(R/N&E) will be low or inscrutable.

C. The Conditionalization Problem

According to EAAN, (a) P(R/N&E) is low or inscrutable, and (b) the naturalist who sees the truth of (a) has a (Humean) defeater for R. But why think (b)? There are plenty of other propositions the naturalist N believes, and on some of them, the probability of R is high: for example, *N is an American and R is true for most Americans*. Suppose N believes this proposition: does that deliver him from the proposed defeat? Which beliefs of his are such that if R is improbable on *them*, then he has a defeater for R? This is a deep question, going to the heart of the nature of defeaters and defeat; it is also a crucial question for EAAN. Richard Otte has excellent things to say about the conditionalization problem, and part of what I say here can be thought of as a de-

velopment or application of some of what he says. I'm not quite convinced by his own solution to this problem, however; I'll say why when I discuss his chapter.

Perhaps we can most profitably enter the conditionalization problem by considering a certain kind of objection to EAAN, one that isn't directly represented by the essays in this volume, but has been raised by Carl Ginet and John Perry.[23] According to Carl Ginet,

> if we delete this component (the difference between theism *simpliciter* and austere theism) and consider just the hypothesis T– that there is a perfect being who creates everything else, then it looks as if we could argue in just the same way Plantinga argues concerning P(R/N&E&C)[24] to the dismal conclusion that P(R/T–&E&C) is low or unknown. Now how is it that the theist is allowed to build into her metaphysical hypothesis something that entails R or a high probability of R but the naturalist isn't? Why isn't it just as reasonable for the naturalist to take it as one of the tenets of naturalism that our cognitive systems are on the whole reliable (especially since it seems to be in our nature to have it as a basic belief)?[25]

What preserves R from defeat, for the theist, Ginet suggests, is the fact that the theist accepts not just austere theism but theism *simpliciter*. Theism, of course, entails but is not entailed by austere theism (as Ginet thinks of it).[26] Hence theism can be described as T– (austere theism) plus a little something extra: T–$^+$, as we might put it. Why can't the naturalist do the same thing? Why can't he move to N$^+$ by adding a little something, perhaps the proposition that "our cognitive systems are on the whole reliable" (Ginet) or "we have won the evolutionary lottery"[27] or a general proposition O "to the effect that the initial conditions of the development of organic life and the sum total of evolutionary processes (including ones as yet unknown or only dimly understood) were and are such as to render P(R/N&E&C&O) rather high?"[28] P(R/N$^+$&E), naturally enough, is not low or inscrutable. Doesn't this look like a good way to stave off defeat?

Well, maybe, but the maneuver certainly has a nasty, suspicious look to it.

[23] Ginet, "Comments on Plantinga's Work"; John Perry (in personal discussion). The objection is referred to at least obliquely by Alston's essay in this volume (as well as by Fitelson and Sober, "Plantinga's Probability Arguments").

[24] Here Ginet is thinking of EAAN as I gave it in *WPF*, where the probability at issue was the slightly more complex P(R/N&E&C), C a proposition setting out some of the main features of our cognitive system (see *WPF*, 220). That additional complexity, I think, is unnecessary.

[25] Ginet, "Comments on Plantinga's Work," 407.

[26] If, as some think, all the properties ascribed to God in theism are entailed by the property of being a perfect being, theism and austere theism would be equivalent. Accordingly, perhaps we should define austere theism instead as the view that we human beings have been created by a very powerful and knowledgeable being.

[27] John Perry, in discussion.

[28] O'Connor, "An Evolutionary Argument against Naturalism?" 535. O'Connor also follows *WPF* in adding the unnecessary C.

Still, why *can't* the naturalist move to N*, thus staving off defeat of R? Better, perhaps he has believed N* all along, in which case there is no question of *moving* to it. He points out that part of his position is and has been that we have won the lottery. Granted: it is unlikely, given our evolutionary origins, that our cognitive faculties would be reliable; but the unlikely often happens, and, fortunately for us, it happened here. He concedes that the probability of R on N&E is low or inscrutable; but he adds that he also accepts N*, on which of course the probability of R is 1. Thus, he says, he has a defeater-defeater for the defeater provided by N&E and the Probability Thesis. Better, he says that the fact that he has *all along* accepted N* means that *N&E & p(R/N&E) is low or inscrutable* isn't a defeater for R in the first place.

What shall we say of this move, besides that it is excessively slick? What precisely does it come to? There are really *three* maneuvers here. Ginet suggests that the naturalist take R itself as part of naturalism, in which case he can point out triumphantly that R is entailed by naturalism so understood, and hence not at all unlikely with respect to it or its conjunction with E. That's the first maneuver; the second is Perry's suggestion that upon seeing the bearing of N&E on R, the naturalist should just add to his noetic structure the proposition L that we have won the lottery, pointing out that R is not improbable or inscrutable with respect to N&E&L. And the third maneuver, O'Connor's, is to add to one's noetic structure the proposition that there is a true proposition P such that R is probable with respect to N&E&P.

The first thing to see is that these procedures can't be right *in general;* if they were, every defeater could automatically be defeated.[29] Suppose I believe the Bible is a special revelation from God and is therefore infallible: everything it affirms is true. Sadly enough, however, I read Mark Twain and H. L. Mencken in an insufficiently critical frame of mind and come to believe U, the proposition that the Bible is unreliable and full of egregious errors. (I form the opinion that a proposition's being affirmed in the Bible confers no more probability upon it than its being affirmed in any other ancient text.) U, I should think, is a defeater for any proposition I accept just on the basis of Biblical teaching. But now consider some belief *B* I do hold just on that basis—perhaps the proposition that Jesus Christ turned water into wine at the wedding feast at Cana. U looks like a defeater for *B*. Could I defend *B* from defeat just by adding a little something to U? Ginet suggests that the naturalist add R itself to naturalism; could I analogously add *B* to U, thus moving to U*, pointing out that the probability of *B* on U* (i.e., U&*B*) is neither low nor inscrutable? Or better, since this *adding* something isn't really relevant, could I point out that I believe not merely U but *B*&U, adding that this conjunction entails *B*, and claiming triumphantly that I therefore do not have a defeater for *B*? As Quine says in another connection, that is not the method of true philosophy.

[29] As Otte in effect points out.

Perry's suggestion is slightly different; what he thinks the naturalist should add is L, the proposition that we have won the evolutionary lottery. The implied scenario is this: the naturalist comes to see that naturalism and the Probability Thesis together provide a defeater for R, but then responds by adding L to his noetic structure, thus acquiring a defeater-defeater, a defeater for his defeater of R. R is unlikely with respect to N&E, he concedes, but now he also believes L and N&E&L; and with respect to *that*, again, R is neither improbable nor inscrutable. But this can't be right either. Consider the probabilistic argument from evil against theism and consider the analogue of Perry's response: "Well, I concede that the existence of God is unlikely given all the suffering the world displays, but I have a defeater for this defeater. I believe that we have won the divinity lottery, and, despite its improbability, that there is indeed such a person as God." Again, not the method of true philosophy. If you discover that you have a defeater *D* for one of your beliefs *B*, you can't in general deliver *B* from defeat just by noting that you believe the conjunction of *D* with some other proposition *D** such that *D&D** entails *B*. In particular, if you believe naturalism and the Probability Thesis and see that this threatens a defeater for R, you can't preserve the latter from defeat just by noting that you also believe N&E&L, with respect to which the probability of R is 1.

O'Connor's suggestion is a bit different again; it is that the naturalist should add

> There is some true proposition *P* such that P(R/N&E&*P*) is high

to his noetic structure. But clearly this is no better than the two preceding suggestions. It would be like conceding that the existence of evil is a defeater for theistic belief, but suggesting that this defeater can be defeated by adding that you think there is some other true proposition *P* (theism itself perhaps?) such that the probability of theism with respect to *P* together with evil is high. Once more, not the method of true philosophy.

What is at issue here, clearly, is this problem of conditionalization. But how can we state the problem more exactly? The dialectical situation goes as follows: I propose that naturalism and the Probability Thesis together offer a defeater for R. The naturalist then replies:

> Right, *P(R/N&E) is low or inscrutable* is indeed true; but there are plenty of *other* propositions *Q* I believe (for example, that R holds for the vast majority of Americans, and I am an American) such that P(R/N&E&*Q*) is *not* low or inscrutable. Why shouldn't I conditionalize on N&E&*Q* instead of just on N&E? Indeed, you yourself, in estimating P(R/N&E) appeal to other beliefs, such as current scientific beliefs as to how motor output originates. In conditionalizing just on N&E, you are neglecting some of the available evidence; there are plenty of propositions I believe such that R is probable with respect to their conjunctions

> with N&E. But then you haven't really given any reason at all for thinking that I have a defeater for R in naturalism and the Probability Thesis.

Thus the objector: she points out that there are propositions Q such that (1) the naturalist believes Q, and (2) the probability of R on $Q\&N\&E$ is *high*: and the principle of total evidence (or perhaps something else in the neighborhood), she says, implies that this fact delivers R from defeat by *N&E & P(R/N&E) is low or inscrutable.*

Now it is certainly true that there are plenty of propositions Q meeting those conditions: for example, there is R itself. But does this deliver R from defeat, for the naturalist? That's the question. More generally, the question of conditionalization is this: which beliefs B are such that if P(R/N&E&B) is high, then *N&E & P(R/N&E) is low or inscrutable* does not constitute a defeater, for the naturalist, of R? Which beliefs are *defeater-deflectors*[30] with respect to the conjunction of naturalism with the Probability Thesis? One answer is that *all* such beliefs deliver R from defeat; if there is any proposition meeting those two conditions, then the naturalist does not have a defeater for R in naturalism and the Probability Thesis. But of course *this* line of thought implies that it is simply impossible to get a probabilistic defeater for R or indeed for any proposition P that I believe. For suppose I believe P: Then for any proposition Q that I come to believe, there will be a proposition that meets the analogue of the above two conditions: P itself. So on the current suggestion, I don't acquire a (probabilistic) defeater for B, no matter what I come to believe. But that can't be right; surely there can be probabilistic defeaters. I believe on the usual grounds that the sphygmomanometer I have just purchased is at least reasonably reliable; I then learn that it was made in a factory owned by a wealthy eccentric whose aim it is to discredit the medical establishment and that 90 percent of the sphygmomanometers made in his factory are unreliable. Then I have a defeater, and a probabilistic defeater, for my initial belief that the instrument in question is reliable. And can't I also acquire a probabilistic defeater for R in my own case? I come to believe that I am a brain in a vat, or a victim of a Cartesian evil genius, or have ingested XX, that anti-reliability drug, and also that 90 percent of those thus afflicted are no longer cognitively reliable. Then don't I have a defeater—a Humean rationality defeater—for R?

So which beliefs meeting (1) and (2) *are* defeater-deflectors for R? That is a difficult question indeed; I certainly don't know how to give a complete and rigorous (or even a complete and unrigorous) answer to it. But of course it isn't necessary to have such an answer in order to see that one does in fact

[30] Not defeater-defeaters. The latter would require that one first have a defeater D for R, and then acquire another belief that defeats D. A defeater-deflector, on the other hand, prevents D from being a defeater in the first place. A proposition is a defeater-deflector for a given belief B *with respect to a specific proposition* (and relative to a noetic structure): in this case, *N&E & P(R/N&E) is low or inscrutable.*

have a probabilistic defeater in a given case. (Indeed, presumably the way to try to answer this general question is by considering a lot of particular cases.) Further, something a bit more general can be said. First, neither R itself nor any proposition equivalent to it (e.g., *(R v (2 + 1 = 4)) & – (2 + 1 = 4))* is a defeater-deflector here. Second, conjunctions of R with other propositions *P* the naturalist believes—e.g., *(2 + 1 = 3) & R*—will not be defeater-deflectors, unless *P* itself is; more generally, propositions *P* that entail R will not be defeater-deflectors, unless a result of deleting R from *P* is a defeater-deflector.[31] Third, no proposition *P* that is evidentially dependent upon R for *S*—i.e., such that *S* believes *P* on the evidential basis of R—is a defeater-deflector for R. Thus *either R or naturalism is true* is evidentially dependent, for me, upon R with respect to me, as is *either R or Friesland is larger than the United States* and O'Connor's *there is some true proposition P such that P(R/N&E&P) is high*. It would be nice to be able to say more, and to see more deeply into this situation.

II. Replies to Individual Authors

It is now time to reply to each of the essays individually. The length of my reply is not proportionate to my estimate of the importance or cogency either of the essay or of my reply; instead, my aim is to reply to those points that seem of most interest or significance with respect to EAAN. I have tried to reply to the essays in an order that will promote economy: in particular I order my response in such a way that later replies can conveniently refer to earlier ones.

A. To Talbott

I turn first to William Talbott's seminal contribution—seminal for me personally, because in an earlier incarnation it led me to see something important about defeaters. This is (a) that there are what I called above *purely alethic* rationality defeaters as well as proper-function rationality defeaters, and (b) that there are *Humean* rationality defeaters.[32] Now Talbott believes EAAN is stronger if taken *locally* rather than globally, that is, if it calls into question, not one's entire epistemic establishment, but just the reliability of certain segments of the whole. In his version, therefore, the naturalist acquires a defeater for those beliefs of his that are produced by his "abstract faculties"—the ones responsible for theoretical mathematics, logic, science, and metaphysics. Since the belief that naturalism is true is produced by those fac-

[31] Where *P* entails R, a result of deleting R from *P* will be any proposition *Q* such that *Q* is logically independent of R and such that *P* is logically equivalent to the conjunction of R with *Q*.

[32] See section 1A above. A good bit of what I have to say in reply to Talbott is to be found in that section.

ulties, he gets a defeater for N, but not for everything he believes. Specifying the argument in this way, says Talbott, gives me the conclusion I seek, but strengthens the argument by weakening its premises a bit.[33] (Sadly enough, he thinks the argument thus strengthened still fails.) Second, my premise was that P(R/N&E) is *low or inscrutable;* Talbott is prepared to concede that P(R/N&E), R specified to those abstract faculties, is *low*—or at any rate he holds that he has a good objection to my argument even given that concession.

How, precisely, does Talbott propose to object to EAAN? First, he proposes analogous cases—cases where one believes something *A* that is *relevantly* analogous to N&E, believes that P(R/*A*) is low, but isn't required by rationality to reject R. For example, suppose I believe that I have ingested H, a drug that produces serious hallucinations in 95 percent of those who ingest it: so P(R/I've ingested H) = .05. But suppose I also believe that 5 percent of the population has a gene which produces a protein blocking the effects of H, and that I am in that 5 percent. Then, says Talbott, I do not have a defeater for R. That is because a certain requirement of total evidence is in force: "there is general agreement that a narrower reference class based on more evidence . . . takes priority over a broader reference class based on less evidence." R is unlikely with respect to one bit of my total (relevant) evidence, but likely with respect to another segment of it, a segment which for present purposes takes priority over it. And the idea, presumably, is that the naturalist has or may have additional evidence that tells *in favor of* R, evidence such that R is probable with respect to the conjunction of that evidence with N&E.

I think Talbott is right specifically about H. This is in part because the effects of H are limited to perception; so presumably I remember thinking I had the blocking gene (and remember my grounds for so thinking) before ingesting H. Perhaps I also notice no change in the operation of my perceptual faculties. If so, the rational belief here is that I am not suffering from hallucinations. So Talbott is right about H. It is far from clear, however, that he would also be right about H' (in a similar situation), where H' attacks reliability generally, not just perceptual reliability.

Now I argued that one can't rationally employ epistemically circular arguments in order to assure oneself of R. (My "epistemically circular" is Talbott's "pragmatically circular.") Talbott, however, believes epistemically circular arguments *are* acceptable, in this context, and that a proposition *A* such that the argument

A
therefore
R

[33] "If Plantinga's argument fails in the first [local] case, it must also fail in the second [global]."

is epistemically circular, can nonetheless be part of total evidence, and also such that it can rationally be conditionalized on in this context. To argue this case, he proposes "Example 2: first example of the Tardy Revelation": BT believes he has taken H, and also that 5 percent of the population has a gene that blocks the effects of the drug, but has no belief as to whether or not he is among the 5 percent. It then seems to him that his doctor calls and tells him that he does indeed display that gene. Under these conditions, Talbott says, the rational thing for BT to believe is that the doctor has indeed called, that BT does indeed possess the blocking gene, and that he is still reliable. The example is designed to show that it is sometimes rational to rely on belief or evidence produced by the very faculty (in this case perception) whose reliability is in question. Total evidence can thus sometimes include such evidence, contrary to my strictures on epistemically circular reasoning.

But here I think Talbott is mistaken. First, an important observation: to consider this example properly, we must make an amendment. That is because Talbott is assuming, in this case, that I have no reason to question the accuracy of my *memory*, and can properly rely on it—for example, for such information as whether my perceptual faculties are working in a way different from the way they worked before I ingested the drug: "he has to admit that he has not noticed much change in his perceptions. He wonders if the effects of the drug only manifest themselves gradually." This introduces a substantial disanalogy with EAAN, where, of course, there is no question of *change* in the operation and reliability of the relevant faculties, and thus no question of being able to detect such change by way of memory or anything else. This leads me to suspect that Talbott's friendly offer to strengthen the argument by putting it locally rather than globally is something of a Trojan horse: it is far from clear that if the argument doesn't work when applied just to perception, it will not work for faculties generally. So suppose we construct an example like Example 2 where it is the reliability of all of one's faculties that is in question (just as in EAAN), being careful to exclude the suggestion that I can sensibly think I could detect changes in the operation of my cognitive faculties, if there were change.

Suppose, therefore, that I take a good dose of XX, which induces not merely perceptual unreliability but global cognitive unreliability. I believe that 95 percent of those in this condition are no longer reliable; I also believe that 5 percent of the population has the blocking gene; but I have no belief as to whether I myself have that gene. I then have a defeater, so I say, for R. Now suppose I come to believe that my physician has telephoned me and told me that I am among the lucky 5 percent whose reliability is unimpaired by ingesting XX. Do I now have a defeater-defeater? Or do I still have a defeater for R? To put the example in Cartesian terms, suppose I believe that I am a victim of a Cartesian evil demon 95 percent of whose victims (as I believe) are cognitively unreliable; I then come to believe that this Cartesian demon her-

self tells me that I am among the lucky 5 percent whose faculties remain reliable, and form the belief that my faculties are in fact reliable. Am I rational in so doing? Now the fact is (I think) that if I were in either of these situations I would believe the doctor or the demon, would believe or assume R, and would carry on just as before the whole miserable scenario began. Furthermore, this is just what (so I think) cognitive proper function requires; if at first I had a defeater for R, I now have a defeater-defeater for that defeater. That is, I have a *proper-function* defeater-defeater. What proper function would require, in the situation in question, is that I believe or assume that I am reliable.

But here we must note that distinction between proper-function rationality defeaters and purely alethic rationality defeaters; we must also note the existence of Humean rationality defeaters. In this situation, either I don't acquire a proper-function defeater for R at all—or, if I initially do, I acquire a proper-function defeater-defeater for that defeater when I come to believe that the doctor or the demon has given me the good news. But I do have a purely alethic rationality defeater here, and I also have a Humean defeater. Perhaps we can see this more clearly if we shift to a third-person perspective. I have this second cousin Sam whom I have never actually met; I naturally assume that Sam's cognitive faculties are reliable. But then I learn that Sam has ingested XX and that 95 percent of those who ingest XX are cognitively unreliable; I now have a defeater for my initial assumption that Sam is reliable. Next, however, I learn that Sam believes his doctor has telephoned him to tell him that he is among the lucky 5 percent who have that blocking gene, thus remaining reliable. In learning that he believes this, would I now have a defeater-defeater for my original defeater for the belief that Sam is reliable? I don't think so. I would properly reason as follows: there is only a 5 percent chance that Sam is reliable. True, he now thinks his doctor has called him to tell him he has that blocking gene; but very likely his cognitive faculties are not functioning reliably in producing that belief. I would continue to think it unlikely that he is reliable and, of course, would not believe that he *is* reliable.

Shouldn't the very same reasoning apply in my own case (with the added complication that I would also appropriately mistrust this very reasoning, as well as any other in which I might engage)? I think so. In this situation I should be in precisely the Humean condition noted above. Perhaps I would be unable, as a matter of fact, to do other than assume R; but when in a quiet moment I reflect on the fact that 95 percent of those who, like me, have ingested XX are unreliable, and that there is thus a high probability that my believing my doctor has told me that good news is itself a product of unreliable cognitive functioning, I will not know what to think. I realize that I am unable to give up R or other ordinary beliefs; still, I also realize that there is only a slim chance that my beliefs are for the most part true. I am in the Humean condition. This condition, it seems to me, is just what proper function dictates in the circumstances of the present example. And this condition is also

what proper function dictates for the naturalist who sees that P(R/N&E) is low or inscrutable.

A couple of final comments. First, Talbott thinks my account of defeat is defective: "On Plantinga's account, rationality exhibits a kind of rigid dependency on the past. Judgments that one's cognitive faculties were probably unreliable would be irreversible, if the only evidence one might obtain of their reliability required the use of those very faculties." As I see it, this is partly right and partly wrong. As to what's partly wrong: I certainly don't think that once you have a defeater for R, it is not possible rationally to get back in the original condition of accepting or assuming R. I suffer from a delusion, thinking I've been captured and envatted by Alpha-Centaurian cognitive scientists; I then have a defeater for R. But I recover from this deplorable condition, recognizing it for the delusion it was; I no longer think I am a brain in a vat; now I accept R in perfect rationality. Giving up the view that I am a brain in a vat can be perfectly rational, and once I lose my defeater for R, I will once again and in perfect rationality assume or accept R.

Of course (and this is the part that's right) I can't, so to speak, *reason* my way out of having a defeater for R. When I have the defeater, I no longer accept R, but then I also have a defeater for any steps of reasoning I might take by way of getting out of this condition; I should be as doubtful about any of them as about R itself. Thus I can't rationally move to the condition of no longer having a defeater for R by way of acquiring evidence for R; for any evidence I might think I acquire is such that I have a defeater for it, in having a defeater for R. There is no sensible strategy for changing belief[34] such that by following that strategy I can move from the condition of not believing R to the condition of believing it. But that is no detriment to rationality.[35]

Finally, according to Talbott:

> when our cognitive faculties, individually or collectively, judge themselves to be reliable on the basis of the total evidence available, and there is no other basis for judging them to be unreliable, I think it is a mistake to claim that that result is objectionably circular. Indeed, if one considers the alternatives (that they judge themselves to be unreliable or that they make no judgment about their reliability), judging themselves to be reliable provides the strongest kind of reason for trusting them.

Here again there seems to me to be something partly right and partly wrong. You ordinarily take it for granted that your faculties do in fact function reliably; even if you are doubtful, on a given occasion, about some specific bit of cognitive functioning, you assume that the rest of your cognitive functioning (e.g., the doubt itself, the whole cognitive background) is proceeding properly. Here there is nothing objectionably circular, epistemically or otherwise.

[34] Such as updating belief by conditionalizing, or Jeffrey conditionalizing.

[35] See Bas van Fraassen, "Conditionalization, a New Argument For," *Topoi* 18 (1999): 93–96.

This state of mind is natural, normal, and perfectly proper; it is precisely what proper cognitive functioning calls for. But the making of this assumption does not function as a *reason* for R. It is not that you believe R on the *evidential basis* of this assumption, as if you believed R on the evidential basis of itself. And now suppose you come, one day, to entertain a global *doubt* as to whether your faculties are reliable. That doubt, I think, cannot be rationally resolved or assuaged by reasoning or argument; for, of course, the doubt attaches with the very same strength to the premises involved in the reasoning, as well as to the thought that if those premises are true, then so is the conclusion. Hence the fact that my faculties "judge themselves to be reliable" does not give me a good reason to believe that they *are* reliable. There is no way to reason oneself out of such a predicament; here salvation will have to be by grace rather than works.

B. To Bergmann

Like Talbott, Michael Bergmann is prepared to "grant to Plantinga the conclusion of the first stage of his argument—that P(R/N&E) is low or inscrutable when specified to us." Grant happily accepted. But then Bergmann goes on to produce a powerful and penetrating argument against the *second* stage of my argument, according to which the naturalist who recognizes this has a defeater for R. He begins by noting that Thomas Reid, his hero and mine, holds that R, the belief that one's cognitive faculties are reliable, is a "First Principle." It is a deeply ingrained belief or assumption that isn't accepted on the basis of other propositions, and that, accepted in this basic way, has a great deal of warrant. Out of these Reidian materials Bergmann constructs a response to EAAN:

> But the commonsense naturalist can respond as follows: "Even if a naturalist believed that P(R/N&E) is low or inscrutable, this needn't give her a defeater for R. For she could have *non*propositional evidence for R that is sufficiently strong to make belief in R rational, reasonable, and warranted—even for someone whose total relevant *propositional* evidence, k, was such that P(R/k) is low or inscrutable."
>
> What matters for my Reidian response to Plantinga is that *we believe R noninferentially on the basis of some sort of nonpropositional evidence and thereby have a lot of justification or warrant for it.*

Now I agree with the second quote: we do indeed, just as Reid said, believe R noninferentially, and this belief, held in that basic way, has a good deal of warrant for us. But I disagree with the first quote: the fact (as Bergmann and I see it) specified in the second quote does not deliver R from defeat for the reflective naturalist. Allow me to explain.

But first, note that Bergmann concedes that naturalism and the Probability Thesis would be a defeater for our beliefs about *other* creatures: "It seems that the conjunction of the belief that N&E and the belief that P(R/N&E) is low or inscrutable *does*—when specified to these hypothetical creatures—constitute a defeater for our belief that *their* cognitive faculties are reliable." Why, then, don't they constitute a defeater for R for the naturalist when R is specified to himself? Because we all, naturalist and theist alike, are constructed in such a way that we naturally accept or believe R in the basic way; more important, so taken, it has a great deal of warrant for us, enough to overwhelm the defeating tendencies of naturalism and the Probability Thesis. When the naturalist sees the truth of the Probability Thesis, what (properly) happens is only that she realizes that one thing she believes—namely R—is unlikely on another thing she believes—N&E.

So far Bergmann. By way of getting on with the explanation I owe, the first and crucial thing to see here, I believe, is part of the topic of section 1A above: the difference between *proper-function rationality* and *warrant*, and the correlative distinction between proper-function rationality defeaters and purely alethic defeaters. As I concede in that section, proper function requires that the naturalist continue in assuming R, despite her seeing that P(R/N&E) is low or inscrutable. Proper function would require the same of someone who believes that she has ingested that reliability-destroying drug XX and that P(R/I've ingested XX) is very low, indeed even zero. But it doesn't follow either that R has warrant for her under those conditions, or that she doesn't have a defeater for it. True, she doesn't have a *proper-function* rationality defeater for it; but she does have (so I claim) a *purely alethic* rationality defeater for it, and consequently, if she is reflective, also a Humean defeater.

I've said most of what I have to say about these distinctions in section IA; here I specify it to Bergmann's contentions and add a bit of detail. So first, it is quite true that the naturalist who accepts the Probability Thesis will undoubtedly continue to assume R, conducting her cognitive life under that assumption; and in so doing, she will display nothing by way of proper-function irrationality. The same would go for me if I believed I'd ingested XX and that all who ingest it become unreliable. I would still step aside if it seemed to me a large truck was bearing down on me, and I would still try to avoid falling (unroped) long distances when rock climbing. Furthermore, this is just what proper function requires; our design plan, one thinks, calls for just this sort of cognitive behavior under that sort of condition. That is because disbelieving or doubting or withholding R would have disastrous cognitive consequences.

Still, however, I *do* have a defeater of *some* sort for R, if I believe that I display some condition *c* such that R is false with respect to all who display that condition—or if I believe that I display a condition *c* such that R is false with respect to most of those who display that condition, and I know of nothing relevant that distinguishes my condition from theirs. This is so even if proper function requires continuing to believe R in these circumstances. What I

have is a *purely alethic* rationality defeater. Proper function requires that the belief that R persist in my noetic structure, but only because of the presence and power of processes not aimed directly at the production of true belief. These processes are aimed instead at the production of belief with some other virtue—in this case, the virtue of enabling me to carry on an acceptable cognitive life even though I believe what would, but for the activity of these processes, be a proper-function rationality defeater for R. I won't repeat the definition of a purely alethic rationality defeater; but the quick and dirty idea is that a person has such a defeater for a belief *B* provided she is such that if only the truth-aimed processes actually involved in the production and sustenance of *B* were operative, she would have an ordinary proper-function rationality defeater for R. Further, upon reflecting on her condition she will come to have a Humean defeater. She believes that she has taken this drug and that most or all of those who take it are such that their cognitive faculties are no longer reliable; she knows of nothing to distinguish her condition from theirs. True: she has that powerful inclination to believe or assume R despite her belief that she's taken XX; but of course she can't sensibly take *that* as evidence for her being reliable: presumably she would have that inclination whether or not the drug has worked its unhappy effect on her.

But then doesn't the same go for the naturalist who realizes the truth of the Probability Thesis? Doesn't he too have a Humean defeater? He believes that N and E are true, and that given that condition, the probability of R is low or inscrutable. He also realizes that in the case of some *other* group of creatures (as Bergmann concedes), this would lead him to withhold the proposition that their cognitive faculties are reliable. True, he also has this strong impulse to continue to believe or assume R. But (as in the above case of the person who believes she's ingested XX) he can't take that as evidence for his reliability; no doubt he would have that inclination to believe R whether or not it is true. When, in quiet moments away from the Sturm und Drang of everyday life, he reflects on his cognitive situation, he will come to withhold R—realizing all the while that he can't withhold R under the conditions of ordinary life or stop forming beliefs under the assumption that it is true. When he reflectively considers his situation, he will withhold R, joining Hume in that cognitively desolate condition.

But couldn't he reason about himself in just the way Bergmann reasons about him? Bergmann suggests or points out that R has much warrant for us, that it has it in the basic way, and that we therefore have nonpropositional evidence for R, just as we do for perceptual and memory beliefs. I agree; I think Bergmann is right. But can the *reflective naturalist* also believe this? He notes, of course, that he has that strong inclination to believe R; but can he, in this circumstance, take the strength of that inclination as measuring the degree of warrant R has for him? I think not. That is because R has warrant for him only if it is produced by cognitive faculties functioning properly and successfully aimed at the production of true belief. But of course he can't sensibly

suppose that his cognitive faculties are successfully aimed at the production of true belief. For he realizes that they are *successfully* aimed at the production of true belief only if they are in fact reliable; and of course that is precisely the question at issue. So he can't sensibly take it that his cognitive faculties are successfully aimed at truth. Further, he also realizes that if N&E is true, there is no reason to think his cognitive faculties are aimed at the production of true belief at all, whether successfully or not. (For example, if his beliefs supervene on neurophysiological processes plus environmental condition, they are presumably not aimed at anything at all; they are more like what Stephen Jay Gould calls a spandrel.) Therefore *he* can't suppose that his beliefs have warrant. And hence he can't sensibly follow Bergmann in thinking that he has nonpropositional evidence for R—evidence strong enough to overwhelm the defeating tendencies of his belief in the truth of N&E and the Probability Thesis.

Finally, Bergmann slyly proposes to assimilate my argument to a general or typical skeptical argument. This is an astute move and a relevant consideration: if my argument is really just a specific version of a general skeptical argument, then it presents no special problem to the naturalist. But how does the proposed assimilation proceed? The typical skeptical argument, says Bergmann, goes as follows:

> X: I currently have *these* experiences and beliefs
> Y: P(R/I currently have these experiences and beliefs) is inscrutable
> Therefore
> It is reasonable to doubt R.

In considering Y, of course, we are to abstract from the *truth* of those beliefs; we are to consider the probability of R just on my *having* those beliefs, together with my having those experiences. But this argument doesn't seem to me at all like EAAN. That is because the conjunction of X and Y isn't any kind of defeater at all for R. Suppose God created angels who are especially impressive epistemic agents: they know far more than we, and make many fewer mistakes. Would R be probable with respect to their having the experiences and beliefs they do have (remembering to abstract from the truth of those beliefs)? There is no reason to think so. They could be magnificent cognitive agents whether or not R was probable with respect to their current beliefs and experiences. No doubt R would be probable with respect to something *else* about them—God's intentions in creating them, or the way in which their beliefs vary with respect to certain kinds of input—but there is no reason to think R (with respect to them) would be probable with respect to their beliefs and experiences. In the same vein, suppose we learn of Sam, not that he has ingested XX, but that the probability of R specified to him is inscrutable on his current beliefs and experiences. Would that give us a defeater for R with respect to him? Surely not; that fact isn't really relevant to R.

Bergmann, to be sure, seems to disagree: "But my replacement for N&E in the typical skeptical objection example—namely, that I currently have these experiences and beliefs—is also relevant because whatever it is that I have going for my belief that R is going to be among my beliefs and experiences (or at least many people think that this is so)." Perhaps it is true that whatever it is that I have going for my belief that R is going to be among my beliefs and experiences; at any rate I have no wish to dispute it. But even if it *is* true, it doesn't follow that if R has warrant for me, then R will be probable with respect to my beliefs and experiences. If, for R to be warranted for me, I had to *infer* it from my present experiences, then R will have warrant for me only if it is probable with respect to those experiences. But clearly that is not how warrant works. My memory beliefs (let's assume) have warrant for me; but not by way of those beliefs' being probable with respect just to my current experience and belief.

I conclude, therefore, that X and Y do not together constitute any kind of defeater at all for R, and that there is little similarity between EAAN and this skeptical argument.

C. To O'Connor

Tim O'Connor sees that one can't properly evaluate EAAN apart from a theory of warrant (or at least rationality), and a view as to the origin and provenance of human cognitive faculties. As for the first, he adopts for present purposes the view I set out in the first two chapters of *Warrant and Proper Function*, according to which "a belief has warrant, for a person, if it is produced by her cognitive faculties functioning properly in a congenial epistemic environment according to a design plan successfully aimed at truth." He then observes that in some contexts it is perfectly rational to (continue to) hold a given belief even if you have powerful evidence against it (the Case of the Purloined Letter). In this way (on this view) it may be perfectly rational to continue to believe in God, despite the apparent counterevidence of evil. This could happen as follows: perhaps we have been designed in such a way that proper function requires one who encounters and reflects on this alleged evidence to continue to take it for granted that (of course) there is such a person as God—even if she hasn't the faintest idea why he allows evil.

Now O'Connor suggests the same should go for the naturalist who sees that P(R/N&E) is low or inscrutable. "An astute naturalist," he says, "will note first of all that the argument doesn't, as a psychological matter, tend to push people to cognitive despair." It is plausible to think, therefore, that cognitive proper function requires retaining R as an assumption, even in the face of evidence against it. "We're designed to start off accepting R without evidence and to continue on that way."

This last seems entirely sensible and indeed no more than the truth. Still, it doesn't imply that the naturalist (or the person who thinks she's a victim of a

Cartesian evil demon, or a brain in a vat, or that she's taken XX) cannot get a defeater for R.[36] In particular, it does not imply that she cannot acquire a Humean defeater for R. In section 1A above and in the replies to Talbott and Bergmann I've already said most of what needs to be said here. But by way of specific application to O'Connor's contention: notice that what he points out suggests only that the naturalist should take it that *proper function* requires that one continue to assume R, even in the face of his acceptance of the Probability Thesis—that is, the proposition that P(R/N&E) is low or inscrutable. But first, that doesn't show that R has *warrant* for one under those circumstances.[37] Under those circumstances, R has proper-function rationality, for me, just as does my confidence that I will recover from a usually fatal disease—even though the statistics I know strongly suggest otherwise. In this case *the optimistic overrider* comes into play; the purpose of the bit of cognitive functioning sustaining that confidence is not that of producing true belief, but of belief conducive to survival. This belief, then, will display proper-function rationality; but it will not display warrant. And the same may be true in the case of R. It may be that this tendency to accept or assume R, even in the face of evidence against it, has as its purpose not the direct acquisition of true belief but rather something else—fending off the cognitive disaster that would result from seriously giving up R in the context of everyday life.

But then, as we saw, the fact that proper function requires continuing to assume R in the face of evidence against it does not protect R from defeat. Reflecting on my belief that E&N, I see that the ultimate purpose of my entire cognitive establishment is to promote fitness, not the acquiring of true beliefs: "The principle chore of the nervous system," as Patricia Churchland says, "is to get the body parts where they should be in order that the organism may survive. . . . Truth, whatever that is, definitely takes the hindmost."[38] Its ultimate purpose is not to promote true beliefs; and have I any reason at all for thinking that its proximate purpose is to promote true beliefs? None that I can see. Furthermore, reflection on EAAN shows that P(R/N&E) is at best inscrutable. Once I see and reflect on this, I will acquire a Humean defeater for R, and thus for my other beliefs as well. In reflective moods, I simply won't know what to believe; I realize that ordinarily I can't help forming belief in the ordinary way, which includes the assumption that R is true (and the fact is I have a defeater for that belief too); but I also believe that this is just an imposition upon me by my nature, an imposition such that the probability of its being connected with truth is not high.

So the fact O'Connor points to—that proper function requires continuing to accept R in the face of the belief that N&E and the Probability Thesis—doesn't at all deliver the naturalist from defeat of R; it delivers her from a

[36] Here, see section 1A above and the reply to Talbott.
[37] See the reply to Bergmann.
[38] "Epistemology in the Age of Neuroscience," *Journal of Philosophy* 84 (October 1987): 548.

proper-function rationality defeater but not from either a *purely alethic* rationality defeater, nor from a *Humean* defeater.[39] And the same distinction between proper-function defeaters and Humean defeaters also calls into question O'Connor's next claim that "Strictly speaking, there *couldn't* be a defeater for R, for any creatures in any possible world." Why not? Because, he says,

> that would require a *design plan* which reliably aimed at truth in some circumstances, a part of which was that we ought to give up belief in R when we take note of certain of our beliefs about the world and about related conditional probabilities. But as that potential upshot would lead one to abandon the attempt to form true beliefs, it surely could not be part of a design plan with that very end.

The first thing to see, I think, is that a design plan, even a cognitive design plan, need not be such that all or even any of its modules are aimed at the production of true belief. (Of course, if none of them is thus aimed, the beliefs produced will not have *warrant* in the above sense.) But let's agree initially that *our* design plan *is*, in fact, reliably aimed at truth, and let's agree further, as O'Connor thinks, that our design plan requires that we not give up R, not even in the face of evidence against it. What follows, again, is that we couldn't acquire a *proper-function* defeater for R; it doesn't follow that we couldn't acquire a Humean defeater for it. Perhaps, under ordinary conditions, we can't help assuming R. Perhaps this tendency on our part is not just directly aimed at the production of true belief but also *indirectly aimed* at it: cognitive disaster would ensue upon seriously doubting or disbelieving R in the circumstances of everyday life. Still, we humans might be so designed that in reflective moments we can come to believe that we are in a condition (having ingested XX, being a brain in a vat, and, so I say, N&E) with respect to which R is unlikely; we can see further that the sensible thing to believe with respect to *someone else* in that condition is that she is not reliable; and we can see, finally, that there is nothing we know about ourselves that relevantly distinguishes us from her. Upon reflection on our epistemic condition, we will then come to doubt R with respect to ourselves; and this doubt is in accord with the design plan. (Note that a part of our design plan directly aimed at truth can conflict with a part indirectly aimed at truth; in certain circumstances the former can tend to lead us to doubt R while the latter tends to lead us to maintain it.)

Finally, in a move reminiscent of Bergmann, O'Connor proposes to "bring the issue into sharper focus" by "posing a dilemma" for me: in effect, his claim here is that the naturalist can make the same reply to EAAN as the theist can to the claim that the probability of theism on evil is low. He proposes

[39] See above for the distinction between these kinds of defeaters. Of course O'Connor can't properly be held responsible for failing to note the distinction between proper function defeaters, purely alethic defeaters, and Humean rationality defeaters; I failed to make that distinction in setting out the argument in *WPF* and "Naturalism Defeated."

four possibilities as to how the theist should respond to this claim (given that she accepts it); the first of these responses is clearly inappropriate; in the case of each of the others, he asks, Why can't the naturalist respond in that same way to his acceptance of naturalism and The Probability Thesis?

Now how, exactly, from the point of view of theism and 'proper functionalism' *should* the theist respond to the belief that P(T/E), the probability of theism on evil, is low or inscrutable? Well, first, she might question its truth. For presumably the idea is that P(T/E) is low because it is thought that (1) probably God would not want to permit evil and would be fully capable of preventing it; hence (2) it is likely that he would create a world in which there is no evil. But is (1) even reasonably obvious? God considers all the possible worlds: he proposes, of course, to create a really good one. Now what are the good-making characteristics in the case of worlds: what makes one possible world better than others? No doubt it is hard to give a general answer; but certainly one good-making characteristic would be divine *incarnation* and *atonement*. In worlds with this characteristic, the second member of the Trinity endures unimaginable suffering and even the pain of separation from God himself ("My God, my God, why have you forsaken me?")[40] in order to redeem guilty creatures who have turned their backs upon God—what could match this for good-making? Indeed, wouldn't all the best worlds contain incarnation and atonement? If so, however, then God would choose for actualization[41] a world that contains incarnation and atonement; but any such world would also contain sin and evil. So the existence of evil is only to be expected, if theism is true, in which case it is hard to see why P(T/E) should be rated low.[42]

But suppose the Christian theist goes so far as to agree that P(T/E) is indeed low; *then* how should she respond to this belief? By continuing to maintain that belief, I should say, because the alleged fact in question is certainly not a defeater for Christian belief,[43] and hence not for theistic belief. First, the theist will not have in E a defeater for T if she is in possession of other propositional evidence for T that is stronger than that provided by E against it. But suppose the theist goes still further yet, conceding for purposes of argument what by my lights is quite false: that E is propositional evidence *against* T and that there is no propositional evidence *for* T. Even so, of course, it doesn't follow that E is a defeater for T; that is the moral of the Case of the Purloined Letter O'Connor cites. In that case, propositional evidence is overwhelmed by the evidence of memory. In this case, the theist will quite properly think that God has provided him with cognitive faculties—for example,

[40] Matt. 27:46.

[41] *Weak* actualization: see Plantinga, *The Nature of Necessity* (Oxford: Clarendon Press, 1974), 173 ff.

[42] This is (of course) just the basic story line of this sort of theodicy: for fuller development see Plantinga, "Supralapsarianism, or O Felix Culpa," in *Faith and Evil*, ed. Peter Van Inwagen (Grand Rapids: Eerdman's, forthcoming).

[43] See Plantinga, *Warranted Christian Belief* (Oxford: Oxford University Press, 1999), chap. 14.

the Sensus Divinitatis and the Internal Instigation of the Holy Spirit[44]—that afford nonpropositional evidence for the existence of God; this evidence overwhelms the propositional evidence from evil against T. The theist will quite properly think this is why he has no inclination to reject theism in the face of the (alleged) evidence offered by evil.

"But if the theist should say this, why should not the naturalist say the same concerning the 'evidence' coming from his own tenacious belief that R?" I think we have already seen why not. The reflective naturalist notes that indeed he does have a tenacious tendency to accept R; but he cannot sensibly say that this tenacious belief that R is evidence, nonpropositional or otherwise, for R. The ideas that a memory belief and the "evidence of the senses" constitute *evidence*, carry with them a double presupposition. This presupposition is that first, the bit of the design plan governing the production of such beliefs—memory beliefs and perceptual beliefs—has as its purpose the production of true beliefs, and second, that as a matter of fact memory and perceptual processes are reliable and hence do produce mainly true belief. But the reflective naturalist will doubt both of these, as well as the corresponding presupposition with respect to R. For it isn't likely that the ultimate purpose (if any) of his cognitive faculties is that of producing true beliefs; and there seems no particular reason to think that it is a proximate purpose either. If, for example, content supervenes on neurophysiological and other physical properties, the content of his beliefs will really be no more than unintended by-products. Natural selection selects for adaptive, fitness-inducing cognitive mechanisms which, at a certain degree of complexity, carry with them content; but then content is no more selected for than is the color of blood.[45] If so, it is not the case that the mechanisms producing his beliefs are aimed at truth. But second, as we have seen above P(R/N&E) is low or inscrutable; so he will also doubt that his faculties produce mostly true beliefs. Therefore both conjuncts of the above presupposition are doubtful, for him, *in excelsis*. Hence he cannot sensibly respond to naturalism and the Probability Thesis in the way the theist properly responds to the belief that E & P(T/E) is low—even given the gratuitous assumption that P(T/E) *is* low.

D. To Van Cleve

James Van Cleve and Ernest Sosa set their remarks in a more general context involving skepticism and the Cartesian circle: this is illuminating. So is the distinction between levels of knowledge—a distinction they themselves make but also attribute to Descartes: animal knowledge vs. reflective knowledge (Sosa), *cognitio* vs. *scientia* (Descartes). It's worth noting, however, that the conclusion of EAAN is not just that the naturalist doesn't *know* R or N&E

[44] See *Warranted Christian Belief*, chaps. 6, 8.
[45] See the text at note 22 above.

(either by way of cognitio or scientia) but that he has a rationality defeater for it—a Humean rationality defeater.

1. *The perspiration objection.* Van Cleve's interest, he says, is in the premise of EAAN according to which if the naturalist has a defeater for *R*, then he has a defeater for *everything* he believes. In this connection he considers the "perspiration objection" I raised in "Naturalism Defeated": P(the function of perspiration is to cool the body/N&E) is low or inscrutable, as is P(Holland, Michigan, is about thirty miles from Grand Rapids/N&E); hence on the principles underlying EAAN, N&E is a defeater for those two beliefs; but surely it isn't; so there is something seriously amiss with the principles underlying the argument. This objection apparently ascribes the following premise ('Principle 1' as Van Cleve calls it) to EAAN:

> (P1) For any propositions A and B that I believe, if B is improbable or inscrutable with respect to A . . . , then A is a defeater for B.

Now Van Cleve notes that I reject (P1) and thus (naturally enough) do not rely on it for the conclusion that

> (1) N&E & (P(R/N&E) is low or inscrutable)

is a defeater for R. He suggests instead that what I must be relying on is

> (P2) For any propositions A and B that I believe, if B is improbable or inscrutable with respect to A, then A is a defeater for B *unless* A derives its warrant from B.

He then points out that (P2) makes a lot of trouble for everyone, not merely those who accept N&E.

I don't doubt that it does, but I am not now nor have I ever been a supporter of (P2): that is because it is clearly false. Most contingent propositions I believe—for example, that at the moment there is a large robin in my backyard—are, I should think, such that their probabilities with respect to, say, the Fundamental Theorem of Calculus, are either low or inscrutable; furthermore, their warrant, for me, does not derive from the warrant, for me, of this theorem. Yet surely that low or inscrutable probability doesn't constitute even a ghost of a defeater for them.

(P2) is unacceptable; but then why *do* I think (1) is in fact a defeater for R? Van Cleve apparently believes that I can properly think this only if I am prepared to subscribe to some principle along the lines of (P1) or (P2): "To avert the havoc Principle 1 threatens to cause for our knowledge, we must *either* reject it outright (in which case we do not have a basis for saying that N&E is even initially a defeater for R) *or* . . ." Well, I don't know of any general prin-

ciple here that will do the job. (Finding such general principles is always a pretty tall order.) But why would I need such a principle? Indeed, wouldn't the right procedure for here be broadly inductive; shouldn't we, at least in this specific sort of case, follow Chisholm, in being particularists?

I would argue that (1) is a defeater for R by way of *analogies*. N&E (if true) gives us crucial information about the origin, provenance, and development of the cognitive faculties in question: the process by which they arose, whether they were designed (and if so, by whom), what their purpose is if they have a purpose, and the like. Now analogous cases where the relevant probabilities are low do offer a defeater. Thus if I learn that my new sphygmomanometer was manufactured in a factory such that the probability of its being reliable given that it was manufactured there is low or inscrutable, and have no other relevant evidence, then I have a defeater for my initial assumption that the instrument is reliable. If I come to think that Sam has ingested that reliability-destroying substance XX, or that he is a victim of a Cartesian demon, or that he has mad cow disease, and if I know nothing else relevant about him and also believe that the relevant probability is low or inscrutable, I will have a defeater for my initial belief that R holds with respect to Sam. The same goes, I say, if I come to think I myself am in one of these conditions—or rather, something similar goes: I acquire a Humean rationality defeater for R. And the same goes, I believe, for (1). If a principle is wanted, I'd suggest starting with something pretty limited, something about beliefs specifying the origin and provenance of cognitive faculties. The antecedent of such a principle would capture what is in common to situations in which (a) I believe that I have taken XX, or am a brain in a vat, or a victim of a Cartesian evil genius, and (b) I also believe that the probability of R on the condition in question is low or inscrutable; the consequent would say that in those situations I have a Humean rationality defeater for R. And the claim would be that the antecedent of such a principle applies to the naturalist who accepts the Probability Thesis.

2. *Conditionalization.* Van Cleve next suggests, alarmingly, that perhaps the naturalist can properly respond to EAAN by conditionalizing on R itself: "Why especially may he not do so if R has high intrinsic warrant for everyone?" Since I have little to add to what I said about this topic in section IC and in the replies to O'Connor and Bergmann, I'll just ask the reader to consult those passages.

3. *Is a defeater for R really a defeater for everything else I believe?* Van Cleve goes on to raise another very interesting question: why should we suppose that (as I claim) having a (Humean) defeater for R gives me a (Humean) defeater for everything *else* I believe? It might do so, he says, if "whenever we accept any deliverance of a faculty, our doing so rests on the implicit premise that that faculty is reliable." But of course (as he points out) this *implicit*

premise thesis is not plausible: I might accept the deliverances of perception, for instance, without ever having so much as raised the question whether it is reliable, and certainly without giving myself an implicit argument one premise of which is that the faculty in question *is* reliable. This implicit premise thesis, however, is not why I think you get a Humean defeater for all that you believe if you get a Humean defeater for R. In order for your acceptance of memory beliefs to be warranted, it is not necessary that you actually believe that your memory is in fact reliable. Notice, however, that if you come to think it is *not* reliable (in a certain range, say) *then* you have a defeater for any memory belief (in that range), provided there is no other source of warrant, for you, for the belief in question. Similarly: you read Freud in an injudicious frame of mind (after a couple of martinis, say); you come to think theistic belief is a product of wish fulfillment and also that the probability of a belief's being true, given that it is a product of wish fulfillment, is low. You then have a defeater for the proposition that the source of theistic belief is reliable; and that in turn, gives you a defeater for theistic belief itself. In general, if you have considered the question whether a given source of information or belief is reliable, and have an undefeated defeater for the belief that it is, then you have a defeater for any belief such that you think it originates (solely) from that source. But then the same applies in the present context; in acquiring an undefeated defeater for R, I get a defeater for all of my beliefs.

4. Epistemically circular arguments. One final matter: in this same connection as well as earlier in his chapter, Van Cleve (like Talbott) raises the question whether I might not be able to base an argument for the reliability of a certain faculty, perhaps over a certain range of its application, on beliefs that themselves are produced by that faculty:

> Is there anything objectionably circular about such arguments? I think not. . . . The really damaging kind of circularity, the kind to be shunned, is the kind in which not merely the *truth* of an argument's conclusion, but *knowledge* of it, is a necessary precondition of obtaining knowledge of the argument's premises. The Cartesian argument sketched above is not circular in that way, for to repeat: clear and distinct perception gives us knowledge (if only at the cognitio level) even if we do not yet know that clear and distinct perception is reliable.

Van Cleve's suggestion is that the same goes with respect to EAAN: why can't the naturalist concede that P(R/N&E) is low, but then give an epistemically circular argument for the conclusion that R holds anyway?

Now it is at least arguably sensible to hold, in *some* cases, that one can base an argument for the reliability of a certain faculty on premises (beliefs) that themselves are a product of that faculty. Perhaps I have never considered the question whether my perceptual faculties are reliable (this question has so far never occurred to me); I immediately conclude that they are, on the grounds

that, as far as I can remember, they have produced beliefs of which the vast majority were true. (There were those times—the first time I saw a bent stick in water, the first time I saw a North Dakota road in summer with that deceptive appearance of wetness—when they misled me; but those occasions have been few and far between.) This is plausibly a sensible procedure.

According to Thomas Reid, however, "If a man's honesty were called into question, it would be ridiculous to refer to the man's own word, whether he be honest or not. The same absurdity there is in attempting to prove, by any kind of reasoning, probable or demonstrative, that our reason is not fallacious, since the very point in question is, whether reasoning may be trusted."[46] Isn't that equally sensible? Suppose I am worried about whether R is true with respect to me: I seem to remember ingesting XX a couple hours ago, and believe that nine out of ten people who have done so suffer from serious cognitive disability within the next couple of hours, so that R no longer holds with respect to them. Now suppose I try to reassure myself by an epistemically circular argument, perhaps an argument of the following form:

I believe p_1 and it's true,
I believe p_2 and it's true,
.
.
.
I believe p_n and it's true;

so most of my beliefs are indeed true; so my faculties are reliable. Is this a reasonable procedure? Surely not; here Reid's point holds.

There is an important distinction between these two kinds of cases. In the kind of case cited by Van Cleve (or that I cited in support of Van Cleve) there has been no reason to question R and in fact the question whether R is true hasn't even come up. Belief in R, furthermore, is plausibly thought to be properly basic; it has considerable warrant in the basic way. Under those conditions, reasoning in the suggested way is acceptable; in any case, I don't propose to dispute it. But it doesn't follow *once one has a question as to whether R* (you've taken XX or whatever), that the same kind of reasoning is acceptable; and indeed it isn't. If you have thought about whether R is true and have a doubt as to whether it is, you can't properly assuage that doubt by giving yourself an argument; for you will, or will if proceeding rationally, have the same doubt about the truth of the premises of the argument, as well as your belief that if the premises are true, so is the conclusion.

In the EAAN case then: if you accept naturalism and the Probability Thesis and as a result come to doubt or wonder whether R is true, you cannot sensibly lay that doubt to rest by way of an epistemically circular argument.

[46] *Essays on the Intellectual Powers of Man*, in *Inquiry and Essays* (Indianapolis: Hackett, 1983), 276.

E. To Sosa

Ernest Sosa's chapter is typically illuminating: he sets his discussion in the context of his own epistemology, as well as reflection on ancient skepticism and Descartes's epistemological efforts. In the process he raises at least four important issues.

1. *A defeater for everything?* After granting for the sake of argument that R is defeated by 'DERIVATION':[47]

N&E & (P(R/N&E) is low or inscrutable)

Sosa asks why this should mean that the rest of the naturalist's beliefs are also defeated. A child who has never thought about R need not be irrational in holding beliefs produced by her cognitive faculties (i.e., in holding beliefs); why then should there be any problem with withholding R but also accepting various everyday beliefs? "Even granted we accept DERIVATION, and even granted this requires us to withhold on R, why should this lack of belief about one's faculties make it irrational to uphold our simple beliefs about food and shelter, if a similar lack does not make such beliefs irrational in a child?" Sosa's point is this: the child does not believe R but nonetheless rationally believes many other propositions; therefore it is possible to withhold R but nonetheless rationally believe many other propositions. So why can't it be that I have a defeater for R all right, but do not have a defeater for the rest of my beliefs?

No doubt a child or for that matter a full-grown adult can rationally believe much without believing R, if only because she has never thought about R one way or the other; perhaps she doesn't so much as have the concept of reliability. (Even so, though, if she *does* have that concept, doesn't she do something like *assume* R? Or perhaps *implicitly* believe it, just as I implicitly believed, a moment ago, that 348 + 1 = 349, even though I had never entertained that thought?) Similarly a naturalist who has never raised the question of reliability, or seen its connection with the Probability Thesis (that *P(R/N&E)* is low or inscrutable) will not have a defeater for R or anything else he believes. But of course, a *reflective* naturalist can't thus protect her beliefs from defeat—particularly if she sees what Sosa concedes: that naturalism and the Probability Thesis give her a defeater for R. And won't she have a defeater—in particular, won't she have a Humean rationality defeater? Return once more to that reliability-inhibiting substance XX, and suppose I believe that (1) I have ingested XX, (2) P(R/I've ingested XX) is low or inscrutable, and (3) this gives me a defeater for R; and suppose as a result I withhold R. Can I then sensibly think I don't have a defeater for the *other* beliefs I hold, on the grounds that it is

[47] Actually Sosa's DERIVATION is a bit broader; but the difference doesn't matter for present purposes.

possible (for a child or an ingenuous naturalist) to hold these beliefs rationally without believing R? I fail to see how. If I have thought about and reflected on the question whether R is true with respect to me, and as a result *withhold* R, and if I see the connection between R and the likelihood that the other beliefs I hold are true, then it seems clear that I have a Humean rationality defeater for those beliefs as well. Perhaps I can't help holding them, but when in reflective moments I think about my epistemic situation and consider the beliefs with which I find myself, I will no longer accept them—except, perhaps in a sort of ironic and ambiguous way, the way in which Hume holds everyday beliefs.

2. *What's so special about R?* Sosa notes next that the naturalist apparently knows *many* propositions that are improbable with respect to N&E—for example, that this particular stone is round, or that I have just drawn the ace of spades. She does this just by taking a look at the stone (or card) and seeing that it is round (or the ace of spades). So why, he asks, can't she do the same thing with respect to R? What's so special about R? He goes on to join Van Cleve in speculating that perhaps the "basis for the conclusion or for the assumption that R specifically is thus susceptible" is the *implicit premise thesis*, according to which "a deliverance of a faculty when accepted is based on an implicit reliance on the reliability of the faculty, and that this reliance takes the form of an assumption that plays the role of an implicit premise in reasoning whose conclusion is the acceptance of the deliverance"; like Van Cleve, he points out that there is little to be said for this thesis.

Here my reply is the same as to Van Cleve. First, I certainly don't accept the implicit premise thesis; there are many beliefs I hold which, as far as I can see, are not accepted on the basis of reasoning of any sort, explicit or implicit. Many perceptual beliefs are like that, as are many memory beliefs and many a priori beliefs (for example, those that are self-evident); and I see no need to postulate some implicit (unconscious?) reasoning in which I do go through such a step of inference. Well then, why can't I simply note that R holds for me, just as I perceive that the stone is smooth and round and that I have just drawn the ace of spades, despite the improbability of these propositions on N&E? True, R is unlikely on N&E, but then so are the propositions that this stone is smooth and round and that I've just drawn the ace of spaces. If that fact doesn't give the naturalists a defeater in the latter cases, why think it does in the former?

This is really a form of the objection Bergmann urges: doesn't the naturalist have nonpropositional evidence for R, just as he does for the propositions that the stone is round and that he has just drawn the ace of spades? My reply to Sosa, therefore, is the same as the answer to Bergmann. First, even though the naturalist does not have a *proper-function* rationality defeater for R, he does have a *purely alethic* defeater for R. We can see this by considering our judgment in the case of someone else—Sam, say—about whom we know just

that he has ingested XX and that the probability of R given that he's ingested XX is low or inscrutable: we would then have a defeater for R with respect to Sam, and would not believe it. But the processes governing the sustenance of R (specified with respect to Sam) in our noetic structure are probably aimed at truth. Hence, if in the analogous case with respect to myself, the only processes involved were truth-aimed, I would have a proper-function rationality defeater for R specified to me. And the same goes when it is N&E and the Probability Thesis that I believe rather than that I've ingested XX.

So first, I say that under these conditions the reflective naturalist has a purely alethic rationality defeater. But second and more important, he also has a Humean defeater for R. He realizes that *N&E & P(R/N&E) is low or inscrutable* would be a defeater for his assumption that R in the case of *other creatures*; he also realizes that he knows nothing relevant to distinguish his case from theirs. True, he has that strong inclination to believe or assume R; but for reasons given above, that inclination isn't a defeater-deflector; it doesn't protect R from defeat. (Just as my belief, if I hold it, that those *other* creatures have a strong inclination to believe R with respect to themselves does not protect that proposition from defeat for me.)

3. Can the naturalist attain reflective knowledge that R? Sosa raises the following question. Granted (which of course I do not grant) that EAAN fails to show that the naturalist has a defeater for his *animal* knowledge (Descartes's cognitio): might it not still be that *N&E & P(R/N&E) is low or inscrutable* defeats his putative *reflective* knowledge (Descartes's scientia)? Here Sosa thinks theist and naturalist are in the same boat. In each case, reflective knowledge or scientia is achieved by producing "an epistemic perspective underwriting the knower's belief from her own epistemic perspective" or "devising an epistemically comforting view of one's universe, according to which one is so constituted and so related to one's world that one's faculties would be not misleading but truth conducive." And here he sees no difference between the naturalist and theist:

> Why can't you as a naturalist develop a view of yourself and your surroundings that shows your situation to be epistemically propitious. You would need to be able to self-attribute ways of acquiring and retaining beliefs that put you so in touch with your surroundings, relative to the relevant fields of interest to you, that you would tend to believe correctly in virtue of using those ways. What precludes your doing so, by means of science, as a naturalist, if the supernaturalist can do so by means of theology?

But here I fear things have gone awry. I'm prepared to agree that the theist needs a story of that sort to attain scientia; but I don't think the naturalist can forestall the defeater that otherwise looms by producing such a story. The problem, as I see it, is not that the naturalist can't think up a story according

to which R holds for us; no doubt he can. The problem is rather that he has a *defeater* for R; and this defeater can't be sidestepped, finessed, defeated or deflected by constructing a story according to which R holds anyway. It is true, of course, that N&E is *consistent* with R, so that in *that* sense N&E is "tenable compatibly with the view that in fact we are quite good cognizers of our surroundings through our use of vision, hearing, etc."; but the one still delivers a defeater to the other. Once more, suppose I believe I am a brain in a vat, or a victim of a Cartesian demon "bent on mischief" as Sosa puts it, or have ingested XX: for definiteness, let's settle on the last. I believe that I have ingested XX and I also believe R is unlikely with respect to this. Could I nonetheless achieve reflective knowledge, scientia, that R is in fact true, with respect to me, by elaborating a story according to which R holds for me even though I've taken XX? Could such a story be either a defeater-defeater or a defeater-deflector in such a way that I wind up with scientia or R? I can't see how.

Perhaps I believe, for example, that while it is certainly unlikely that we human beings should be reliable, given our origins, natural selection has nevertheless inexplicably brought it about that we are. This would be a version of the "Hooray! We've won the evolutionary lottery!" strategy. The problem would be with the status of those beliefs constituting the story. Presumably they would be the sort of beliefs for which evidence is required, if they are to be accepted rationally: but where would we find the evidence? On the other hand, suppose the naturalist just assumes that R is true (despite its unlikelihood with respect to naturalism) and then looks for and embraces the best explanation of its being true. Then his belief in that explanation is *evidentially dependent* on his belief that R, and thus cannot function as a defeater-deflector with respect to R; it runs contrary to the third condition for defeater-deflection. The truth is, so it seems to me, that reflection doesn't give the naturalist scientia or reflective knowledge of R; what it does is give him a reflective (Humean) defeater of R, and hence deprives him even of his animal knowledge of that proposition. How does telling that story Sosa proposes differ from just whistling in the dark?

4. *The worth of reflective knowledge.* I certainly agree with Sosa that reflective knowledge, scientia, is of great value, very much worth having. But *why* is it such a good thing? According to Sosa, in large part because it is in some way *autonomously acquired*: "Reflective acquisition of knowledge is . . . like attaining a prized objective guided by one's own intelligence, information, and deliberation; unreflective acquisition of knowledge is like lucking into some benefit in the dark." "Enlightened discovery is success attributable to the agent; luck in the dark is not." Sosa compares the person who acquires this sort of knowledge to a tennis champion, whose instinctive responses at the net are due to long hours of intelligent practice, so that he can properly take

credit for them as his own doing. When the epistemic agent attains reflective knowledge, his acquisition is "his doing."

Here we may have our doubts—not about the value of reflective knowledge, but about this suggestion as to the source of that value. Sosa's suggestion seems to me unduly to credit human autonomy. Of course it is nothing like postmodern paeans to human autonomy, or even Heidegger's complaint that he can't feel comfortable in a universe he hasn't himself created,[48] but it leans a bit far in that direction. Reflective knowledge is not primarily something the agent—as opposed to God or nature—can take credit for. In doing the reflecting, one is, naturally, relying at every point on 'animal' knowledge—for one's grasp of the principles of logic, for example. One also has to rely on memory, for which one cannot claim credit. Indeed the very capability of forming belief is not something for which we can take credit. A *bit* of what goes on in the acquisition of reflective knowledge is my own contribution, so to say; but of course the vast bulk is not. The same goes for Sosa's tennis champion. She can perhaps attribute a bit of her tennis-playing abilities and excellence to her own efforts: but the bulk of it—being able to swing a racket, see a tennis ball, having arms and legs, vision, and coordination—is clearly nothing she can properly take credit for (unless, in a fit of postmodern excess, she proposes to claim that she has created herself).

F. To Merricks

Trenton Merricks astutely addresses one central premise (the "Crucial Claim") of EAAN:

> (CC) If P(R specified to oneself/N&E) is low or inscrutable, then one who believes N&E has a defeater for R thus specified.

He leaves open the question whether the naturalist, if she is to have a defeater for R, must *believe* that P(R specified to herself/N&E) is low or inscrutable (see his note 4). As I see it, this *is* required for her to have a defeater here; so consider (CC) thus amended. (CC) is indeed a crucial premise of the argument: Merricks believes it is questionable. Why so? He observes first that the *general* claim, according to which

> For any beliefs A and B, if P(A/B) is low or inscrutable, then B is a defeater for A,

[48] See Richard Rorty, *Contingency, Irony and Solidarity* (New York: Cambridge University Press, 1989), 109.

is of course absurdly false; most such cases do not issue in defeat. "Once we see that," says Merricks, "we should also see that anyone who endorses this sort of inference in a particular case owes us an explanation as to why, in that case, the inference is justified." So what is there to be said for (CC), and why should we believe it? I argue for (CC) by way of proposing analogies; Merricks's strategy is to question the analogies.

The first analogy he questions is the one I draw in *Warrant and Proper Function* between that hypothetical nonhuman population and us human beings. Here I think he is right—at any rate, my development of the analogy suffers from a nasty glitch. I begin by arguing that we should think P(R/N&E), R specified to that alien population, is low; so we should think it low with respect to us as well (*WPF*, 228). Fair enough. Then a couple of pages later I claim that one might well instead think that P(R/N&E), R specified to that alien population, is *inscrutable*, rather than low. Again, fair enough. Next, I proposed that the proper attitude on our part toward R (specified to them) would be *agnosticism*. And I then suggested that the same goes for us; one who accepts N&E and thinks P(R/N&E) is low or inscrutable (R specified to us) should be agnostic with respect to R. But Merricks notes a crucial disanalogy. I say that R *with respect to them* is a proposition such that we believe it rationally only if we have (propositional or testimonial) *evidence* for it; but as Merricks points out this doesn't obviously hold for R *with respect to us*. It is far from obvious, he says, that a person rationally believes her own cognitive faculties are reliable only if she believes this on the basis of (propositional) evidence.

Agreed. Not only is that not obvious, it is false; as I see it, the belief that R (specified to oneself) has warrant when it is accepted in the basic way, not on the evidential basis of other propositions. I did not mean to suggest that one needs propositional evidence to be rational in believing R specified to oneself. To construct a proper analogy, therefore (as Merricks recognizes) I should stipulate that we are originally rational (Merricks says 'warranted') in believing R specified to that population (on the grounds, perhaps, that they look a lot like us, live in houses, use tools, have a language, and the like; we then come to believe N&E and the Probability Thesis (R specified to them). And my claim should then have been that this gives us a defeater for R with respect to them.

All right, consider it done. That is how the analogy should go: and I (along with Bergmann: see above) do indeed believe that our coming to believe *N&E & P(R/N&E) is inscrutable* with respect to them gives us a defeater for R with respect to them. Merricks, however, does not; he suggests that in such a case we would *not* have a defeater for R with respect to the population in question, or at any rate *need* not have a defeater for R with respect to them. That is because, he says, the source of our warrant for believing R (specified to them) might be that we have *observed* them:

> Perhaps we should imagine that one has interacted with the members of that population and lived among them. (This seems a standard way to become warranted in believing R specified to some group of persons.) In such a case, it is not at all clear that R specified to the members of that population really *is* subject to a defeater when one comes to believe that they evolved in such a way that the probability of their being reliable, on their so evolving, is low or inscrutable.

Again, agreed. Consider an analogous case: I have been observing Sam for the last five years or so; now I learn that a couple of years ago Sam ingested XX, which in nine out of ten cases induces permanent loss of cognitive reliability. As far as I can tell however, Sam has been a veritable paragon of reliability ever since I first started observing him. My new discovery certainly gives me no defeater for my belief that R holds with respect to him; what I should think instead, as Merricks says, is that he has beaten the odds.

But of course this is not appropriately similar to EAAN, the case at hand. For suppose instead that I come to believe that I *myself* have ingested XX, again, with the supposition that nine out of ten people who do so no longer reliably form beliefs (and thus are no longer correct in believing R with respect to themselves). (Alternatively, suppose I come to believe that I am a brain in a vat or a victim of a Cartesian evil genius, and think that in each case nine out of ten of their victims hold beliefs that are absurdly false.) Here it is clear, I think, that I *do* have a defeater: both a purely alethic defeater and a Humean defeater. True: I won't in fact be able to withhold R under ordinary circumstances. But in reflective moments, when I consider the low probability that I am reliable, given that I've ingested XX, and also realize that I know of nothing relevant to distinguish my case from that of others who have ingested it, I will withhold R.

Now what is the essential difference between this case and the case of Sam? Why don't I get a defeater in the latter case? Obviously because of the evidence provided by my long-standing and close observation of him. But there is no analogue to this evidence to which I can sensibly appeal in my own case. I can't sensibly say, in order to reassure myself about R specified to me, "Look: I have observed myself for a good long time and noted that in fact most of my beliefs have been true, so that I am in fact reliable. I have beaten the odds." This would be precisely the procedure Reid correctly ridicules.[49] I was certainly rational in thinking I was reliable before I ingested XX; but now I've got excellent reason to doubt that I am, and thus excellent reason to withhold the proposition that I am. No doubt I am *unable* (in ordinary circumstances) to withhold that proposition; and no doubt I am perfectly rational, in the proper-function sense, in those circumstances, in continuing to assume R. But I still have a defeater—a Humean rationality defeater.

[49] See the text at note 46 above.

And doesn't the same hold if P(R specified to me/I've ingested XX) is *inscrutable* rather than just low? Suppose all I know is that some who ingest XX become unreliable, but I don't know the proportions: won't I then too have a reason to reject R with respect to them? I will be in the condition of one who initially believes in complete rationality that his new sphygmomanometer is reliable, but then learns that it was produced in that factory owned by the eccentric bent on interfering with modern technology. I know that at least some of the factory's sphygmomanometers are unreliable, and I thereby acquire a defeater for my initial belief that my sphygmomanometer is reliable. Something similar is true in the present case: I acquire a Humean defeater for R.

So suppose I believe I've ingested XX and that P(R/I have ingested XX) is low or inscrutable: then I do indeed have a defeater for R with respect to me. Why does Merricks think otherwise? Because, I believe, he fails to see the dissimilarity between the third-and first-person cases. He suggests that what he calls "The Analogous Claim$_3$"

> (AC_3) If P(R specified to someone else/revised demon hypothesis) is low or inscrutable, then one who believes the revised demon hypothesis about someone else has a defeater for R specified to that other person[50]

is false, as indeed it is. As Merricks point out, you might have had this person under observation for some time and learned that R is true with respect to her. But then he goes on to say that noting the falsehood of (AC_3) gives us reason to think "The Analogous Claim$_2$"

> (AC_2) If P(R specified to *oneself*/revised demon hypothesis) is low or inscrutable, then one who believes the revised demon hypothesis has a defeater for R specified to herself or himself

is *also* false. I say the falsehood of (AC_3) provides no support at all for the claim that (AC_2) is false. The reason is clear: (AC_3) is false because of course you can quite properly *observe* this person and determine by broadly scientific means whether or not he is reliable. Not so in your own case. For, as you realize, your checking to see whether you yourself are in fact reliable is a wholly bootless procedure. It is no more effective than your asking Sam to check himself for reliability (and relying upon his answer), after he has taken XX.

According to Merricks, "once we build p's *initial warrant* into one's noetic structure, I cannot think of a single case in which it is intuitively obvious, or even intuitively plausible, that q is a defeater for p solely on the basis of a low or inscrutable P(p/q)." Here two comments: First, the naturalist can't assume that R *does* have warrant in these cases; for warrant, at least as I see it, requires

[50] Here too we should insert a clause to the effect that the one who believes the revised demon hypothesis also believes that P(R specified to the relevant someone else/revised demon hypothesis) is low or inscrutable.

that the segment of the cognitive establishment producing the belief in question be aimed at the production of *true* belief. If my cognitive establishment has been produced by the undirected processes of current evolutionary science, however, the *ultimate* purpose, at any rate, of my cognitive faculties, if they have a purpose, will be that of producing beliefs that are adaptive, fitness-enhancing, not true beliefs. Furthermore, one can't just assume that their penultimate (or antepenultimate, or . . .) purpose is the production of true beliefs. What the naturalist *can* properly assume, I think, is that R has a great deal of initial *proper-function rationality*: it is another question altogether whether it also has warrant. Still, let's suppose that R does have initial warrant as well as proper-function rationality. Now Merricks says he can't think of any cases where a belief p has initial warrant but is defeated by some belief q by virtue of the fact that $P(p/q)$ is low or inscrutable. But there are many such cases. I'm initially warranted in believing my new sphygmomanometer is accurate (new sphygmomanometers usually are). But then I discover that it has been fabricated in that factory mentioned above, and that the probability of its being reliable, given that it was fabricated there, is low or inscrutable. Now I have a defeater for my original belief. You tell me that last month you climbed the Grand Teton; I form the warranted belief that you did climb the Grand, and admire you even more than I already did. But then I learn from your wife that you have run into psychological difficulties lately and have taken to trying to impress people by claiming, falsely, that you have climbed this or that mountain. Given this belief, the probability that you climbed the Grand is low or inscrutable; and I certainly do have a defeater for it. More to the present point, the assumption that Sam is reliable has warrant for me; but then I learn that yesterday in an experimental mood he ingested XX (and I believe that nine out of ten of those who do that become epistemically unreliable within a couple of hours). If I know nothing further (I *haven't* been observing Sam) then I have a defeater for the assumption that he is reliable, as I do if the probability is inscrutable for me.

G. To Fodor

Jerry Fodor embraces what he calls 'Scientism': the idea "on the one hand, that the goals of scientific inquiry include the discovery of objective empirical truths; and on the other hand, that science has come pretty close to achieving this goal at least from time to time." Now this is a weak and wholly nonstandard use of the term 'scientism'; usually the term includes much more, something like Bertrand Russell's claim that "Whatever knowledge is attainable, must be attained by scientific methods; and what science cannot discover, mankind cannot know."[51] Still, it's a free country; Fodor has a right to use that term any way he wants. What is confusing, though, is that he also suggests

[51] *Religion and Science* (Oxford: Oxford University Press, 1978), 243.

that what *I* call 'naturalism' is the same thing *he* calls 'scientism': "not, at least, if what he [i.e., I] has in mind is an argument against Scientism (or against what he calls 'naturalism,' which, at least for present purposes, we may take to be much the same thing)." This is both false and infelicitous. First, it is false: I accept scientism (as Fodor explains it) but emphatically reject (what I call) naturalism. The former fits in well with theism, and of course modern science emerged from the matrix of theistic (and indeed Christian) belief; the latter, as I characterized it, is obviously incompatible with Christian or theistic belief. Second, it is infelicitous; it confuses matters by suggesting that my argument is directed against what he calls 'scientism'. But when I say that the naturalist has a defeater for R, I certainly don't mean to say that a defeater for R afflicts just anyone who thinks that science aims at the discovery of objectively true beliefs and sometimes succeeds in that aim.

1. *The context.* Now Fodor's main claim about EAAN, as I understand him, is that I am mistaken in thinking that P(R/N&E&C)[52] is low or inscrutable—or perhaps his claim is that I haven't succeeded in *showing* that this probability is low or inscrutable. How, exactly, does this bear on EAAN? Recall that according to EAAN, (a) P(R/N&E) is low or inscrutable, and (b) this gives the naturalist a defeater (a Humean defeater) for R, and hence a (Humean) defeater for anything else he believes, including N&E itself. The argument for (a), in turn, proceeds by way of the truth that

$$P(R/N\&E) = [P(R/N\&E\&C) \times P(C/N\&E)] + [P(R/N\&E\&{-}C) \times P({-}C/N\&E)].$$

Now Fodor's attention is directed entirely at the term

$$P(R/N\&E\&C) \times P(C/N\&E);$$

he doesn't consider the term

$$P(R/N\&E\&{-}C) \times P({-}C/N\&E)$$

at all, apparently on the grounds that the right hand member of that term is negligible: "Plantinga thinks (as I do not) that there is some serious chance it will turn out that we don't act out of the content of our mental states *at all*, a fortiori that the truth of our beliefs can't be what explains our behavioral successes." Fodor thinks there is no serious chance that (my way of putting it) the content of our beliefs does not enter into the causal chain leading to behavior. Right, I agree. Like the rest of us, I'm certain that the content of my beliefs does indeed play a causal role in my behavior; I don't for a minute seriously

[52] Where 'C' names the proposition that the content of a belief enters the causal chain leading to behavior. Here I am taking C to exclude the thought that content supervenes on neurophysiological and other physical properties.

entertain the thought that the content of my belief is causally inert or impotent with respect to my behavior.

But of course that does nothing to show that P(–C/N&E) is negligible. The probability of semantic epiphenomenalism may still be high on N&E (and if so, so much the worse for N&E), no matter how strongly we commonsensically accept its denial; there is no guarantee, after all, that N&E supports common sense. Indeed, the heart of my argument is that it does *not*. And the fact is, as I've already argued, P(–C/N&E) *is* high. That is because naturalism, as I'm thinking of it, includes materialism, or at any rate the idea that human beings are material objects (even if, for example, sets and numbers are not). And the problem is this: it is extremely hard to see, from the perspective of materialism, how it *could* be that the content of a belief, the proposition that is associated with that belief as its content, plays a role in the causation of the behavior. Insofar as a belief enters the causal chain leading to behavior, it is by virtue of its neurophysiological properties, not its content.[53] Attempts to show how it could be that content does enter that causal chain have unanimously come to grief. So I say the probability of P(–C/N&E) is high. But clearly the probability of P(R/N&E&–C) is low.[54] And that means that it doesn't matter all that much to EAAN what P(R/N&E&C) turns out to be. Suppose, for example, P(R/N&E&–C) is .2 and P(–C/N&E) is .8. Then P(C/N&E) will be .2; and then even if we assign P(R/N&E&C) a value of 1, P(R/N&E) will be less than ½ (.36, to be exact)—certainly low enough to be a defeater, for R, for the naturalist. So EAAN emerges unscathed even if we concede everything Fodor argues for.

2. *Fodor's argument.* Still, I did give an argument for the conclusion that P(R/N&E&C) is at best moderately high; Fodor's remarks on this argument seem to me to be wide of the mark. Suppose we briefly explore those remarks. He notes first that my argument depends on the claim that it is quite possible, and perfectly compatible with N&E, that there be creatures with *adaptive* but mainly *false* beliefs. Here Fodor has his doubts:

> Let it be that that *appletree witch is blooming* is false, or lacks a truth value, in virtue of its presupposing that that appletree is a witch. Still, much of what a creature believes in virtue of which it believes that *that appletree witch is blooming* (and in virtue of which the thought that *that appletree witch is blooming* leads to be-

[53] Perhaps Fodor or others might think that content *is* causally efficacious, just because it supervenes (causally or logically) upon neurophysiological properties (together, perhaps, with features of the environment).

[54] We might note that if content supervenes on neurophysiological and other physical properties, then presumably there will be causal generalizations connecting content with behavior; and we might suppose *that* sufficient for content's being causally efficacious with respect to behavior (and thus for the falsehood of semantic epiphenomenalism). Perhaps (though more likely perhaps not); in any event, the probability of R under this possibility is also low.

> havioral successes) are perfectly straightforwardly true. For example: *that's an appletree; that's blooming; that's there; something is blooming; . . .*

This overlooks the fact that, as I constructed the case (and whose case is it, after all?), these creatures form beliefs only of the form 'that *P*-witch has *Q*' for properties *P* and *Q*. (We may add, if we like, that they form general beliefs of the form *all (some) P-witches are Q*, together with propositions appropriately constructible out of these general and singular beliefs.) So the creature in question doesn't believe *that's an appletree* (though he may believe *that witch is an appletree*) or *that's blooming* (though he may believe *that witch is blooming*). Why couldn't there be creatures like that? Not, surely (as Fodor himself notes), because any such creatures would have to believe all the logical consequences (or all the obvious logical consequences) of what he believes; we ourselves do not do that. What his argument needs, he says, is only "the truism that the creature has whatever beliefs the explanations of its behavioral successes require attributing to it." But surely its behavioral successes can be explained as well by virtue of its believing *that witch is blooming* and *that witch is an appletree* as by virtue of its believing *that's blooming* and *that's an appletree*? An explanation of its behavioral successes clearly doesn't require the attribution to it of the sorts of beliefs Fodor mentions.

But Fodor says he's willing to concede for purposes of argument that a set of beliefs could be adaptive though mainly false; he wants to consider the question whether an argument like mine can work even given that concession. He thinks not. *Why* not? Here things get a little murkier. First, he accuses me (and Churchland) of reasoning that (Argument A*) "our behavior could be mostly adaptive even if most of our beliefs were mostly false. So there is reason to think that most of our beliefs *are* false." Churchland will have to speak for herself, but I haven't argued in that fashion since the sixth grade. What I actually said is that there are indefinitely many systems of belief and desire that will yield the same (adaptive) behavior; among these there are indefinitely many where most of the beliefs involved are mostly false; and then, "Here one doesn't quite know what to say about the probability that their cognitive systems would produce mainly true beliefs, but perhaps it would be reasonable to estimate it as somewhat more than ½" (*WPF*, 227). No claim here that we have reason to think most of our beliefs are false; and certainly no argument to that claim from the premise that it is possible. That kind of procedure works only for the ontological argument.

Fodor recaps his argument as follows:

> The polemical situation, according to the opposition, is that a Darwinist has no reason for supposing that evolution picks true believers rather than false-believers-with-cognitive-mechanisms-in-the-charmed-circle. This is because, by assumption, the behavior of the latter would be just as adaptive as the behavior of

> the former. But, I claim, that's not good enough. It can't be plausible that we were selected for the behavioral success of our false beliefs unless it is *independently* plausible (or, anyhow, independently not *im*plausible) that a lot of our behavioral successes have in fact been actions out of false beliefs. But, to repeat, the commonsense view is that, however we may have evolved, our behavioral successes are generally predicated on true beliefs; and that it's generally the truth of the beliefs on which they are predicated that explains the successes of the behaviors; and that, in cases when we have managed to act successfully out of false beliefs, that's mostly been blind luck. As far as I can tell, there is no reason on God's green earth why a proper Darwinism should reject this commonsense view.

Now whatever its precise relevance to my argument, I do in fact believe that "a Darwinist has no reason for supposing that evolution picks true believers rather than false-believers-with-cognitive-mechanisms-in-the-charmed-circle"; this is indeed because "the behavior of the latter would be just as adaptive as the behavior of the former." (More exactly, it is because the probability that evolution would pick true believers as opposed to those lucky false believers, is *inscrutable*, a belief that as a matter of fact I think Fodor shares.)[55] But Fodor (as is sometimes his wont) says poo to that; it's not good enough, he says. Why not? Because, he says, "It can't be plausible that we were selected for the behavioral success of our false beliefs unless it is *independently* plausible (or, anyhow, independently not *im*plausible) that a lot of our behavioral successes have in fact been actions out of false beliefs." But why does this have to be *independently* plausible? In fact, what does it mean for it to *be* independently plausible? Independent of *what*? I'm not sure what he means, but perhaps it's this: if the naturalist has a defeater for R in N&E and the Probability Thesis, it must be because he has something like evidence independent of N&E that as a matter of fact –R is true with respect to him.

But why think a thing like that? I learn that my sphygmomanometer was fabricated by the wealthy eccentric who wants to do in the medical establishment; as far as I can see, the probability of its being reliable, given this origin, is low or inscrutable; I have no other evidence for its reliability. Doesn't that just in itself give me a defeater for my initial assumption that the sphygmomanometer is reliable? Why do I need independent evidence that it is unreliable? Suppose I've ingested that reliability-inhibiting drug XX and I think the probability of R, given that fact, is low or inscrutable. Then don't I have a defeater for R, with respect to me, even if I don't have *independent* evidence that my cognitive faculties are unreliable? Suppose I come to think I am a brain in a vat (and thus, as Fodor says elsewhere, have legitimate grounds for com-

[55] "Since, for all we know, evolution would have chosen false believers if it had been given the chance, the fact that evolution chose *us* isn't, in and of itself, a reason for thinking that we're true believers. There is, as far as I can tell, *no* Darwinian reason for thinking that we're true believers. Or that we aren't."

plaint) and that nine out of ten such unfortunates are unreliable: don't I then have a defeater for R, even if I don't have independent evidence for the conclusion that a lot of my behavioral successes have been actions out of false belief?

But perhaps Fodor has something else in mind: perhaps he thinks it can't be plausible that we were selected for false but adaptive beliefs unless it is independently plausible that a lot of our behavioral successes are out of false beliefs; and by way of proposing that this is *not* independently plausible, he points out that we commonsensically think our successful actions ordinarily arise out of true belief. Indeed we do; but how is that relevant in the context? We ordinarily think we know of lots of cases where we enjoyed behavioral success which arose from true belief (we ordinarily think R holds with respect to us); certainly so, but is that (to use the language of section IC) a defeater-deflector for R? Certainly not; this belief is or is equivalent to a conjunction, one conjunct of which is R itself; hence, according to the second constraint on defeater-deflectors it cannot serve as a defeater-deflector.

I therefore do not believe Fodor identifies serious problems for EAAN, or even for the proposition that P(R/N&E&C) is low or inscrutable.

H. To Ramsey

Bill Talbott is willing to concede that P(R/N&E) is low; Bill Ramsey aims to argue that it isn't—or that there is no reason to think that it is, or at any rate that I have not given good reason to think that it is. How does his argument go?

1. *Truth and belief.* Ramsey notes first that truth is not an intrinsic property of belief (and the property of believing *p* truly is not an intrinsic property of a believer); hence the truth of a belief does not supervene just on the intrinsic properties of the belief (or of the believer).[56] I could be in a certain neurophysiological state *S* and believe *truly* that I see an orange in the bin; but I could have been in that same state and believed *falsely* that there is an orange in the bin (if, for example, there had been a fake orange there). Ramsey then points out that *relational* properties of an organism, properties that are not reducible to (and do not supervene just on) its intrinsic properties, can be adaptive and can be selected for: examples would be good camouflage and the property of being faster than the predators in the neighborhood.

Fair enough; but why does Ramsey point these things out? Because, I think, he believes that EAAN depends on the following thought: properties that are not reducible to (and do not supervene just on) intrinsic neurophysiological properties cannot be selected for. In any event he seems to attribute

[56] "A belief's truth value does not supervene on the belief's intrinsic/neurophysical properties; truthfulness can vary while the neurophysical properties remain the same."

that thought to me in the paragraph that ends with "As Plantinga himself puts it, 'if [truth and falsehood] are not involved in the causal chain leading to behavior, then once more beliefs—or rather, their semantic properties, including truth and falsehood—will be invisible to natural selection.'" But EAAN doesn't depend at all on the thought that relational properties cannot be selected for. Clearly, many can; indeed, fitness itself is a relational property that is not reducible to and does not supervene on intrinsic neurophysiological properties (it depends also on the nature of the environment); and fitness can certainly be selected for.

If that's not what I meant, what *did* I mean? Recall that my argument for the claim that P(R/N&E) is low or inscrutable went as follows: I pointed out that

$$P(R/N\&E) = P(R/N\&E\&C) \times P(C/N\&E) + P(R/N\&E\&{-}C) \times P({-}C/N\&E),$$

where C is the proposition that the *content* of a belief (as well as its neurophysiological properties) enters into the causal chain that leads to behavior, and –C is semantic epiphenomenalism, the proposition that content does not enter that causal chain. (It is only by virtue of its neurophysiological properties, and not by virtue of its content, that a belief makes a causal contribution to behavior.) Then I argued that P(R/N&E&–C) is low or inscrutable and P(–C/N&E) is high or inscrutable; and the quoted sentence was part of the argument for the claim that P(R/N&E&–C) is low (or perhaps inscrutable). I was arguing that the probability of R, *given* semantic epiphenomenalism together with N&C, is low. But the argument doesn't proceed by way of the proposition that if truth and falsehood of belief do not supervene on neurophysiological properties of the believer, then natural selection cannot select for true beliefs.

Rather, the argument goes as follows. Natural selection, of course, operates on *behavior* in a broad sense; it rewards behavior that enhances fitness and punishes behavior that does not. Thus it shapes behavior in the direction of greater fitness. In so doing it also shapes the sorts of structures that *generate* behavior in the direction of greater fitness, greater adaptiveness. Now if content of belief *did* enter the causal chain that leads to behavior—and if true belief caused adaptive behavior (and false belief maladaptive behavior)—then natural selection, by rewarding and punishing adaptive and maladaptive behavior respectively, could shape the mechanisms that produce belief in the direction of greater reliability. There could then be selection pressure for true belief and for reliable belief-producing mechanisms. But under the hypothesis in question, the content of a belief, as opposed to its neurophysiological properties, does not enter into the causal chain leading to behavior. And then it is not the case that a belief produces adaptive behavior by way of being true, or maladaptive behavior by way of being false. So natural selection can't, directly, at any rate, mold belief-producing mechanisms in the direction of the

production of reliability by rewarding adaptive behavior and penalizing maladaptive behavior.

True, both content of belief and adaptive behavior could be effects of a common cause. This would be so if content *causally* supervenes upon neurophysiological properties (together, perhaps, with properties of the environment). As I argue above, in this case too the probability of R will be low (or inscrutable)—as it will if content supervenes *logically* upon neurophysiological and other physical properties.

Furthermore, we can't think of any other way in which natural selection could shape belief-producing mechanisms in the direction of reliability. But then the conclusion to draw is that P(R/N&E&–C) is low (or inscrutable). And this argument doesn't depend at all upon the thought that if truth is a relational property, then natural selection couldn't select for true belief. Indeed, suppose with common sense that semantic epiphenomenalism is *false*: then belief *would* cause behavior by way of its content; but then false belief could cause maladaptive behavior and true belief adaptive behavior, and reliable belief-producing mechanisms could indeed be selected for. What counts here is not whether truth is a relational property (which it is) but whether belief is causally efficacious by way of its content.

2. *Can false belief produce adaptive behavior?* Ramsey suggests next that there is "a perfectly plausible way for truth to be a property that, although irreducible, is nevertheless favored by a process of natural selection." He adds that "On this view, the truthfulness of our beliefs actually helps *explain* their adaptiveness": "If all we want is an explanation of Bob's immediate motor behavior, then we needn't appeal to the truth or falsehood of any of his beliefs. But if we want to know why Bob's behavior proves successful while his cohorts die out, then it clearly does matter that his behavior is generated by true beliefs." Finally, he proposes that "It is, in fact, extremely difficult to give a serious adaptational account of cognitive processes in which truth (or something close to it, like accuracy) does not matter for mental representations."

Now the first thing to see here is that what Ramsey says in these three passages seems to presuppose that the content of belief *is* causally efficacious with respect to behavior. Hence if I am right in thinking P(–C/N&E) much more likely than P(C/N&E), the truth of what he says, if indeed it is true, will not compromise EAAN. The fact is, though, that what he says seems to me false, or more accurately, vitiated by an interesting and important conflation. What's important to see is that if beliefs are in fact (as Ramsey assumes) *representations*, then there are at least two very different kinds of representation. Imagine a running back trying to elude tacklers. He sees the tacklers and the speed and angle with which they are approaching (and the ferocious looks on their faces), but he doesn't ordinarily form beliefs about these matters: there isn't time for that. Nonetheless his cognitive system is guiding his elusive (we hope) behavior. No doubt his cognitive system contains *representations* of

these potential tacklers and their approaching speed and direction; but he need have no *beliefs* on these topics at all. Or think about the mechanisms that regulate one's body temperature. When the temperature of the blood flowing through the hypothalamus exceeds 98.6 °F, nerve cells signal the body that it's time to cool down. Blood vessels near the skin expand to bring more blood to the body's surface, where excess heat can radiate away; sweat glands release perspiration, which cools the body by evaporation. (Evaporation of a medium-sized drop of sweat can reduce the temperature of a quart of blood by about a degree Fahrenheit.) But then there must be something in the hypothalamus like a temperature measuring device, something such that various different states of the device are correlated with different temperatures in the organism's bloodstream.

In the case of a mercury thermometer, what is correlated with different temperatures will be the height of the mercury column; a certain height of the mercury column can thus be said to *indicate* or even 'represent' a certain temperature; and we can speak of a given state of the thermometer as a representation of the current ambient temperature. Still further, we can speak of this representation as having content—*indicator* content, we might say: its content is the temperature it represents or indicates. Further yet, we can also speak of *accuracy* here, although perhaps in a Pickwickian sense: a representation r, of this sort, is accurate if the state that r is in is the one with which r is correlated—in this case, the ambient temperature. Representations of this sort will therefore be *trivially* accurate. In the same way, then, there will be some device in the body that measures temperature, and we can speak of a given state of that device as a representation of the temperature of the blood. This kind of representation—call it an *indicator* representation—is of course very different from belief; it is not that this device, or the organism, or anything or anybody else in the neighborhood *believes* that the temperature is so and so.

Now it is not so much as remotely plausible to suggest that the indicator content of an indicator representation does not (in general) enter the causal chain that leads to behavior. On the contrary: it is because the state in question represents what it does that it causes the behavior that it does; there is an easy explanation of its causing *that* behavior in terms of what it represents. It is because it represents *that* temperature that, when it occurs, it causes shivering. Furthermore, since indicator representations are automatically accurate, natural selection will not be able to select for mechanisms that produce *inaccurate* representations. But none of this, so far, has anything to do with *belief*, or with the *truth* of a belief.

Accordingly, when Ramsey says, "It is, in fact, extremely difficult to give a serious adaptational account of cognitive processes in which truth (or something close to it, like accuracy) does not matter for mental representations," I suspect he is conflating beliefs with indicator representations. At any rate it is the former that are relevant for EAAN, but what he says seems to me clearly

true only of the latter. For it certainly is hard to give a "serious adaptational account" of *indicator representations* in which accuracy does not matter for adaptiveness, but it is easy enough to give a serious adaptational account of *belief* in which truth does not matter for adaptiveness. I gave some suggestions along this line in *Warrant and Proper Function;* Ramsey comments as follows:

> the critical question is whether or not a process of natural selection is likely to generate cognitive mechanisms that are systematically unreliable but somehow prove adaptive. Instead of giving us specific belief-desire pairs that produce life-preserving behavior in certain odd circumstances, Plantinga needs to describe general *mechanisms* or *processes* that would prove adaptive while giving rise to a preponderance of false beliefs.

Fair enough; what has to be shown is how it can be that *false* beliefs can be adaptive, that is, entire *systems* of mainly false beliefs can be adaptive,[57] and adaptive by virtue of their content. Very well, here is one way, specified to the naturalist's way of looking at the world. Suppose naturalism is true and in fact there is no such person as God. Now several naturalists (E. O. Wilson and Michael Ruse, for example) have argued that belief in God, while false, is nonetheless adaptive. So suppose a tribe of cognitively gifted creatures believe that everything (except God Himself) has been created by God; they therefore think everything is a *creature*, something created by God. Suppose further that their only way of referring to the various things in their environment is by way of such definite descriptions as 'the tree creature before me' or 'the tiger creature approaching me'. Suppose still further that all their beliefs are properly expressed by singular sentences whose subjects are definite descriptions expressing properties that entail the property of creaturehood—such sentences as 'The tiger creature approaching me is dangerous' or 'The tree creature before me is full of apple creatures'. Suppose, finally, that their definite descriptions work the way Bertrand Russell thought definite descriptions work: 'The tallest man in Boston is wise', for example, abbreviates 'There is exactly one tallest man in Boston, and it is wise'. Then from the naturalist perspective all their beliefs are false. Yet these can still be adaptive: all they have to do is ascribe the right properties to the right 'creatures'.

There are plenty of other ways in which systems of mainly false beliefs can be adaptive. These tribespeople, for example, might also think that in any situation calling for action, there is just one action which has the nonnatural property of being right; all the others have the nonnatural property of being wrong. Given that there isn't any nonnatural property of being right, their judgments about actions will be for the most part mistaken—though adaptive, if they ascribe nonnatural rightness to fitness-promoting actions and

[57] As Ramsey notes, I try to do this in "Naturalism Defeated."

take the actions they think are nonnaturally right. Or take still another tack: Michael Rea argues that naturalism implies an ontology of gunk, an ontology according to which there really aren't any *objects* (although there is a sort of continuous gunk or goo which may, as they say, be 'propertied' differently in different places). Suppose Rea is right: then since most of our beliefs imply that there are objects, most of our beliefs will be false. Still, our natural way of cutting the world up into objects could be adaptive; if so, most of our beliefs would be false but adaptive.

Now in response to a similar suggestion in "Naturalism Defeated,"[58] Ramsey suggests that "Just because someone thinks that trees are conscious, and thinks of them under the description 'those conscious trees', it clearly doesn't follow that they can't also think, truthfully, that trees grow in the ground"; and he attributes my thinking otherwise to P1 or P2, a couple of clearly false 'principles'. But I don't accept either P1 or P2. Instead, I say that it is certainly possible that there be creatures all of whose beliefs are of the above form, and who are also such that their definite descriptions function Russellianly: in which case all or nearly all of their beliefs, from a naturalistic point of view, would be false (though still adaptive). Still, we can accommodate Ramsey's contrary view here: annex *creaturehood* to the properties expressed by the *predicates* of these speakers instead of to their referring devices. Suppose each (or at any rate a relevant range) of their predicates expresses a property that includes *being a creature*—that is, *being created by God*. Then once more the bulk of their beliefs will be false, from a naturalistic perspective, but can nonetheless be adaptive. And couldn't this easily happen? Ramsey suggests that "this sort of deception would not be easy to pull off, and the mechanisms involved would need to be more sophisticated than those that are straightforwardly truth-tropic"; but I fail to see the problem here. It could easily happen, even from a naturalistic perspective, that a tribe comes to believe that there is such a person as God; and (as many naturalists suggest) this could have adaptive advantages. It could also easily happen that their predicates come to express only properties that entail creaturehood. Where is the problem?[59]

It seems to me, therefore, that one can't sensibly hold that P(R/N&E&C) is high. But what the naturalist must hold, to evade the force of my argument, is *both* that P(R/N&E&C) is high *and* that P(–C/N&E) is low; I say each of these is implausible. The best that the naturalist can do here, I think, is claim that these probabilities are inscrutable. But if I'm right, that also gives him a defeater for R.

[58] A bit of the suggestion from "Naturalism Defeated" is quoted in Fodor's chapter.

[59] It is worth remembering that the question here is whether it is possible that there be creatures like us whose beliefs are for the most part false but nonetheless adaptive. The question is *not*, therefore, whether *we* use definite descriptions in the Russellian way, or whether *our* predicates express properties including *being created by God* (contrary to Ramsey's suggestion in his note 6).

I. To Fales

Evan Fales begins the main part of his argument by rejecting three propositions: (1) the view that there is no causal connection at all between belief and behavior, (2) epiphenomenalism *simpliciter*, the idea that belief does not enter the causal chain that leads to behavior either semantically (by way of content) or syntactically (by way of neurophysiological properties) and (3) the idea that belief is causally efficacious, both semantically and syntactically, but is nonetheless maladaptive. He rejects them all on the grounds that "the neural systems by means of which organisms generate and manage their beliefs are biologically expensive." I don't know whether that is a good reason for rejecting them, but let's not quibble: let's concede that these three are all unlikely on N&E, unlikely enough so that we can ignore them for present purposes.[60] That leaves us with two possibilities: on the one hand semantic epiphenomenalism (–C), the idea that content is causally impotent and beliefs are causally efficacious only by way of their neurophysiological properties, and, on the other, C, the commonsense view, according to which beliefs are causally efficacious both semantically and syntactically, and they (or the mechanisms that produce them) are also adaptive. I believe Fales means to argue that both P(R/N&E&–C) and P(R/N&E&C) are high.

1. *Beliefs and representations.* Consider the first: how does Fales propose to argue that P(R/N&E &–C) is high? What he says is that "a central question is whether it is conceptually possible that semantic content would fail to map onto syntax [neurophysiological properties] so as to preserve truth." This is relevant as follows: if it is conceptually necessary that semantic content map onto syntax in such a way that the content is true, then even if content does not enter the causal chain leading to behavior, that content will nonetheless consist in true or mainly true propositions—so that, contrary to what I argued, P(R/N&E&–C) will be high, and, since the same will be true for P(R/N&E&C), so will P(R/N&E). Now I'm not entirely clear what Fales means when he speaks of semantic content failing "to map onto syntax [neurophysiological properties] so as to preserve truth." What I think he means, however, is that semantic content maps onto neurophysiological properties just if the propositions that get somehow associated with neural structures (or perhaps *adaptive* neurophysiological structures), as their content, are *true* propositions. Further, I believe he means to suggest that this is conceptually necessary. But why believe a thing like that? Why believe that the content of those neurophysiological structures that are beliefs (or adaptive beliefs) is somehow conceptually bound to consist in *true* propositions? That seems at best a pious hope; it doesn't *seem* conceptually necessary at all. How is this

[60] Inspection of the formula on page 9 shows that only a negligible contribution to the value of P(R/N&E) is made by a term involving a P_i such that $P(P_i/N\&E)$ is very low.

supposed to mesh with the fact that many beliefs—adaptive beliefs as well as maladaptive—are *false*? Many of the cognoscenti in this area believe that semantics *supervenes* on syntax (that content supervenes on neurophysiological properties together with certain features of the environment); that surely seems a salient conceptual possibility. At a certain level of neurophysiological complexity, perhaps, content simply appears: a given proposition or group of propositions get associated, somehow, with a given neurological structure. But why suppose the proposition or propositions in question must be *true*? Couldn't they just as well be false? In fact many human beliefs—again, adaptive as well as maladaptive—*are* false; is there some sort of conceptual guarantee that most of them are true? What could that possibly be?

Here Fales makes the following suggestion: "mental representations," he suggests, "get their content in virtue of being caused in the right way by items in the environment." So far, again, this is relatively unproblematic. Actually, there seem to be at least two possibilities here: it could be that content supervenes *causally* on neurophysiology together with environmental items, and it could also be that it supervenes *logically* (in the broadly logical sense) on neurophysiology together with environmental items. It isn't at all clear that either of these or their disjunction is a conceptual truth (whatever precisely that is), but they do seem to be relevant epistemic possibilities. Still, the central question remains: Why think the content, the proposition(s), that thus gets associated with a given neurophysiological structure must be *true*? Couldn't they just as well be false? Is there even the hint of a guarantee that such propositions are obliged, conceptually or otherwise, to be true?

Fales makes still another suggestion: "if a mental representation is caused in the right way by heat, then it's a representation of heat; and if it is not so caused, then it's not a representation of heat." Now of course we are thinking about *beliefs;* so if this remark is apposite, Fales must be thinking of beliefs as among the representations. Perhaps there are representations of various kinds,[61] but among them are beliefs. Well, suppose what he says is so: a mental representation is a representation of x just if it is caused in the right way (whatever that is) by x: why think the representation (belief) in question must be a *true* representation of heat? Suppose a representation (a belief) is caused in me in the right way so that it is a representation of a tree: why suppose it must be a *true* representation of a tree? Can't I have a false belief about a tree? Maybe the tree is a beech, but I think it's an elm: can't that happen? Why does Fales overlook this question?

I venture a tentative suggestion: perhaps he is conflating certain kinds of representations, *indicator* representations, as we may call them, with beliefs. I've already called attention to this possible conflation in my reply to Ramsey; here I can be brief. There are (presumably) internal structures that monitor blood temperature (or sugar or sodium level); a state of such a structure (like

[61] See my reply to Ramsey.

the height of the mercury column in a thermometer) could be said to be a representation of a certain temperature. And perhaps, as Fales says, that state will *be* a representation of that temperature just if it is caused in the right way by the blood's displaying that temperature. If we like, we can say, further, that this representation has a *content*: the state it represents.[62] Still further, if we wish we can even speak (no doubt in a Pickwickian sense) of the *accuracy* of the structure in question: a representation is accurate when its content is the state it displays when the ambient temperature causes it (in the right way) to exhibit a state. (Representations will thus be trivially accurate.) Of course the representation isn't *true* (or false); the mechanism is not speaking truly when it attains the state that matches the ambient temperature, and it isn't speaking falsely either.

But these indicator representations are not beliefs. No one in the neighborhood—me, my body, my blood, the structure, the state of the structure—*believes* that my blood temperature is thus and so.[63] Belief is quite another matter. A belief (typically) represents something or other, but also represents it *as being* this or that. The content of a belief will be a proposition: my belief that the Taj Mahal is pink represents the Taj Mahal, and represents it as being pink. Beliefs, then, as opposed to indicator representations, can be true or false. If my belief represents the Taj Mahal as being, say, the second largest used clothing store in Asia, then the belief represents the Taj Mahal, all right, but *misrepresents* it, and is accordingly false. Now, following Fales, it may be plausible to think that the content of belief and other representations supervenes on neurophysiology (and the environment). It may also be plausible to think that the sort of structure that generates or subtends indicator representations can't fail (given ordinary conditions and proper function) to display the relevant content; the mercury column of a thermometer is bound to achieve the height that goes with the ambient temperature. But it isn't even remotely plausible to think that the content of *belief* is somehow obliged to supervene in such a way as to be *true*. Only magical thinking, I believe, would lead to that conclusion.

"So long as representations are causally linked to the world via the syntactic structures in the brain to which they correspond, this will guarantee that syntax maps onto semantics in a generally truth-preserving way," says Fales. But why should we think so? Suppose my representation of the Taj Mahal as the second largest used clothing store in Asia is causally linked to the world via the syntactic structures in the brain to which it corresponds; why should

[62] A hitch: Given that the content of an indicator state is any state with which it is causally correlated in an appropriate way, its content won't be unique (there will be many states with which it is causally correlated). We can ignore this little glitch for present purposes; in other contexts (e.g., when trying to explain belief content in terms of indicator content) it becomes crucially important.

[63] Well, it could be that *I* believe this; but the representation goes on whether or not I believe it.

that even so much as slyly suggest that this representation must be *true*, that what this belief represents it as being is the way (or a way) it is? *Indicator* representation is perhaps guaranteed to be "accurate" in that extended sense; but the same thing does not hold for belief. Fales gives us no reason at all, so far, for supposing that beliefs are somehow bound to be mostly true by virtue of the way in which they come to have their content. He gives us no reason at all to suppose that P(R/N&E & semantic epiphenomenalism) is high.

2. *False belief and adaptive action.* Fales also faults my way of suggesting that P(R/N&E&C) is not as high as one might think. I said that there are many different systems of belief and desire that would subtend adaptive action; hence it's not particularly probable that creatures whose behavior is adaptive hold mostly true beliefs. But according to Fales, "there are no effective algorithms connecting false belief to appropriate action, as there are when the input is true beliefs and the rules of inference employed are valid or inductively sound." This seems to me hasty. Consider the cognitive agents who think everything is created by God and whose predicates express only properties entailing being created by God. Then, by the naturalist's lights, their beliefs will be mainly false. Still, their beliefs can obviously be adaptive, that is, lead to appropriate action; all that's required is that they ascribe the right properties to the right objects. Thus, for example, if they ascribe the property of being a tiger creature to tigers, and the property of being a dangerous creature to tiger creatures, they will presumably act in appropriate ways. Here there is just the sort of effective algorithm Fales says there isn't (at least if he is right in thinking there are algorithms connecting true belief and adaptive behavior). We can see this as follows. First, say that the *creaturely transform* of a property *P* is a property equivalent to the conjunctive property *P and being a creature;* then take an algorithm that appropriately connects true belief with appropriate action, and replace the relevant properties to which the algorithm directs you with their creaturely transforms. You will then have an algorithm that maps false beliefs onto appropriate action. And the challenge he issues can easily be met, I think (if not for monkeys, then certainly for humanlike cognitive agents) along the same lines.

3. *Tu quoque?* Finally, Fales aims to turn the tables on the theist: if EAAN shows that the naturalist is subject to a (Humean) defeater for R and hence for naturalism itself, then, so he says, a precisely similar argument shows that the theist too has a defeater for R, and hence for theism itself. This is an intriguing maneuver,[64] and leads to fascinating and challenging areas. Here I can't pursue this idea nearly as far as it warrants: the following will have to suffice. There are two separate questions: (a) does the reflective theist (and I'll

[64] A maneuver that (as I note in "Naturalism Defeated," 18 ff.) has also been suggested by Earl Conee, Richard Feldman, Theodore Sider, Stephen Wykstra, and still others.

limit myself here to Christian theists) get a defeater for *R* (taken generally or with respect to specific kinds of belief)? and (b) does the reflective Christian theist get a defeater for *theism*? The first question is the more vexed; given the Christian view of sin, it is not implausible to hold that it adversely affects certain areas of our cognitive establishment, so that in some areas there may be systematic compromise of reliability.[65] I'll address just the second of these. The naturalist, if I am right, gets a defeater for naturalism itself; thus naturalism, so I claim, is self-defeating and hence irrational. Does anything similar apply to the Christian theist?

I think not. A Christian accepts theism, of course, but also the specifically Christian beliefs that distinguish Christians from other theists. Now if Christian belief is true, it is eminently likely that the belief-producing processes that give rise to it are in fact reliable. According to the central teaching of Christianity, God has devised a way in which we human beings, despite the sinful condition into which we have fallen, can achieve a right relationship with God; this scheme involves the life, suffering, death, and resurrection of Jesus Christ, the second person of the trinity. Now God would of course *want* us human beings to know about this scheme, in particular because its efficacy involves, at least in the paradigmatic and central cases, *belief* in it. Thus it is likely, given the essentials of Christian belief, that God would furnish a way in which we human beings can know about this salvific scheme, and hence that the belief-producing processes from which Christian belief emanates are indeed reliable.[66]

Further and more specifically, the Christian typically accepts the authority of the Bible, in particular though not exclusively in matters of faith and morals. But according to the Bible (as both Aquinas and Calvin attest) the *source* of Christian belief, in those who display it, is the Internal Testimony (Calvin) or Instigation (Aquinas) of the Holy Spirit.[67] This belief-producing process or mechanism is reliable, according to Christian belief, because it is designed by God to produce belief in the main lines of the Christian gospel, which, according to Christian belief, is, naturally enough, true. From this narrower perspective too, therefore, we see that, given Christian belief (including belief in the veracity of the main lines of the Bible), it is likely that the belief-producing processes responsible for Christian belief are indeed reliable. But theism is part of Christian belief and entailed by it. Hence with respect to Christian belief the probability that the process(es) producing theistic belief is reliable, is high. But then the Christian theist does not have here a defeater for theistic belief; the *tu quoque* argument fails. Of course it

[65] These areas would have to do with our knowledge of God and our knowledge of other people; sin, according to the Heidelberg Catechism, induces in us a tendency to "hate God and our neighbor," which in turn induces cognitive debility in those areas. This debility, however, is, according to Christian teaching, progressively ameliorated by regeneration and sanctification.

[66] See *Warranted Christian Belief*, 285.

[67] See ibid., chap. 8.

might be that Christian belief is irrational in some *other* way. But it isn't subject to the sort of self-defeat, the sort of self-referential incoherence that afflicts N&E.

J. To Otte

Ric Otte's powerful chapter deals with the problem of conditionalization. What precisely *is* this problem? That in itself is a wholly nontrivial problem, but perhaps we can make just a bit of progress. Consider EAAN: according to Otte, the conditionalization problem is the problem of determining which conditional probabilities should be used in EAAN. This problem arises as follows. According to EAAN, the naturalist who sees that P(R/N&E) is low or inscrutable has a (Humean) defeater for R and hence for everything else he believes. But a naturalist might reply as follows: "Well sure, *that* probability is low, but so what? Why think *that* probability gives me a defeater for anything? There are plenty of other propositions *P* I believe such that P(R/*P*) is *high*. There are even many such that P(R/N&E&*P*) is high. So why think the probability you mention gives me a defeater for R? What's so special about *that* probability?" So there is a problem here: what is the appropriate body of propositions to conditionalize on in this context? According to Otte, "the problem is one of deciding which probability to use in Plantinga's argument"; "the problem," I take it here, is the problem of conditionalization. But perhaps this formulation suffers a bit, as a general statement of the problem; at any rate it would be nice to know what determines which probability is the right one to use in contexts like EAAN. The short answer is that we should use in the argument a probability that works: one that is such that (1) the proposition *P* conditioned on includes or entails N&E, (2) P(R/*P*) is low or inscrutable, and (3) the naturalist's believing *P* together with his seeing the relevant probability, does give him a defeater for R. Another way to put the problem, then: which probability(ies) for R is such that its condition includes N&E and is such that if the naturalist accepts it, and it isn't high, then he has a defeater for R?

Otte's proposal: the relevant probability is P(R/N&E&EXP), where EXP is a proposition stating the naturalist's *experience*. This experience is perceptual but also goes beyond that: Otte thinks there is such a thing as experiencing *that you believe thus and so*. I'm not sure exactly what that would be. There certainly is such a thing as *doxastic* experience—that is, the experience constituted by, or going with, a certain proposition's seeming *true* to one—but is there such a thing as the experience of believing a proposition? This isn't crucial for present purposes, however. For present purposes we could say that EXP includes perceptual and doxastic experience (and any other kind of experience), and what is to be conditionalized on is

N&E&EXP&B,

where B is a proposition stipulating the agent's beliefs, a proposition according to which S believes p_1–p_n, where p_1–p_n is an enumeration of the propositions the subject believes.[68] Otte's claim, then, if I understand him, is that if

(1) P(R/N&E&EXP&B)

is not high, and an agent believes N&E and sees that probability (1) is not high, then that agent has a defeater for R. More cautiously, perhaps, his claim is that this probability is at any rate the best candidate for the post of being the proposition such that if the agent recognizes that R is improbable on it, then she has a defeater for R.

Now I'm not convinced that (1) *is* the best candidate. Otte suggests that I claim that cognitive faculties arising by way of naturalism and evolution would be likely to be unreliable; hence "*all* of our beliefs are called into question. Thus we would be begging the question against his [my] proposal if we were to conditionalize on the truth of any of the beliefs that arise from this source. But that means we cannot conditionalize on the truth of any of our beliefs, since they all arise from this source." We are thinking about whether the naturalist has a defeater for R, and hence whether, if he believes R and N&E and the Probability Thesis, his beliefs are rational. The question is how the naturalist should respond when he is clearly aware of these beliefs and reflects on them. What does rationality require? If, as I claim, he has a Humean defeater, then what rationality requires is that he move to that ironic, reflexively skeptical Humean condition.

But then the issue is, which probabilities are the ones relevant to the question whether he *has* a defeater? Can he sensibly say, for example: "Well, *that* probability is low, but here's one that's high: *P(R/N&E & we've won the evolutionary lottery);* so I don't have a defeater for R after all"? Let's say that he *conditionalizes on* the condition of any probability he uses to determine his propositional attitude toward R. (And of course this determination need not be a result of explicit and conscious reflection.) Then the question is: what does rationality require that he conditionalize on? Well, one thing it requires is that whatever he conditionalize on be among his beliefs; a probability whose condition isn't among his beliefs isn't relevant to his determining his attitude toward R. So if he is to conditionalize on N&E, or N&E&P for some other proposition P, all of these will have to be propositions he *believes*. But then it *seems* that Otte's suggestion has the consequence that the naturalist can't rationally conditionalize on *anything*. And that can't be right, not at least if there is any cogency to EAAN. (It also can't be right, given that in analogous situations (mad cow disease, ingesting XX, being a brain in a vat) one clearly does get a Humean rationality defeater, and gets it by way of perceiving certain probabilities—hence by conditionalization.) It is certainly true, as Otte

[68] Perhaps Otte intends that B be entailed by EXP; if so, we have a harmless redundancy.

points out, that *all* of the naturalist's beliefs are called into question: but that is dialectically posterior to the conditionalization involved. Once one sees the relevant probabilities, *then* one has a defeater for R, and hence for all of one's beliefs, including the relevant probabilities, and indeed for the defeater of R itself. One then has a *universal* defeater, and therefore a defeater that defeats itself.

This can be confusing. In *Warrant and Proper Function* I unwisely followed Hume and Sextus Empiricus in arguing that the naturalist, if rational, will fall into the following sort of diachronic loop: first, he believes N&E and *P(R/N&E) is low or inscrutable* (call that conjunction Q); this gives him a defeater for R and hence for everything else; so he stops believing Q; but then he *loses* his defeater for R and Q; so presumably those beliefs then come flooding back. But then once again he has a defeater for them, and withholds them; and so on, round and round the loop (*WPF*, 235). So what you really have is a loop where N&E keeps getting alternately defeated and reprieved—at t_1 it is defeated, at t_2 undefeated, at t_3 defeated, and so on. I went on to say (even more unwisely) that this situation gives him an ultimately undefeated defeater for R. But this line of thought can't be right. Even if the naturalist got into such a diachronic loop, it wouldn't be the case that he would have an ultimately undefeated defeater for R; what he would have instead is a defeater that is alternately defeated and undefeated, and hence not ultimately defeated—which, of course, is quite different from being ultimately undefeated. And second, why think rationality requires that he get into this loop in the first place? Can't he see in advance what's going to happen?

Let me offer the following correction. The devotee of N&E, if he is canny (or reads what I am about to say) sees that there is a certain *synchronic* structure here. He sees that there is an infinite series of potential defeaters and defeatees; but he needn't himself doggedly plod around any diachronic loop. To see the relevant structure, consider a similar but simpler case: imagine a skeptic who comes to believe –R, that is, that his faculties are not reliable. What he should see (what in any event *we* can see) is that there is then an infinite series of propositions related in an interesting way. At the first level, there is –R, which he believes. But there is a connection between –R and any belief he has, including –R itself: if –R is true, then any belief he has, including –R itself, will be unreliably formed. But *that* belief—the belief that –R is unreliably formed—gives him a potential defeater for –R. Suppose we let '–R(p)' express the proposition that p is unreliably formed. Then at the second level we have –R and –R(–R), which is a defeater of –R. But then at the third level we have –R and –R(–R(–R)), which is a potential defeater of –R(–R), the defeater of –R; and so on. Perhaps we can schematically represent the structure as follows:

level 0: –R
level 1: –R and –R(0) (i.e., –R(–R)) (which is a potential defeater for level 0, i.e., –R)

level 2: –R and –R(1) (which is a potential defeater for level 1 and a potential defeater-defeater for level 0)
level 3: –R and –R(2) which is a potential defeater for level 2 and a potential (defeater-defeater)-defeater for level 0
.
.
.
level n: –R and –R(n-1) (which is a potential defeater for level n-1)
.
.
.

There is an infinite series of propositions here, but not an infinite series of defeaters and defeatees. D is a defeater of B only if D is a defeater of B *for someone;* D is a defeater of B for someone S only if S believes both D and B; but of course no one could believe all the propositions in this infinite series. Still, we might say that each member is a *potential* defeater for the preceding member, in that if someone believed n-1, and saw –R(n-1), he would have a defeater for n-1.

And the important point here is this: in thus acquiring a defeater for level 1, that is, in acquiring a defeater-defeater for the defeater of level 0, one does not lose one's defeater for level 0. As long as the skeptic believes –R, she has a defeater for –R: –R(–R). And she has this defeater even if she also has a defeater for –R(–R). This is extraordinary: in the more usual case, if one acquires a defeater-defeater for a belief B, that is, a defeater for a defeater of B, one no longer has that defeater for B—or else its defeating power is neutralized. But not so here. The difference is that here the original defeatee shows up at every subsequent level. When that happens—when, roughly speaking, every potential defeater in the series is really the potential defeatee plus a bit—the defeater-defeater doesn't nullify the defeater. The defeater gets defeated, all right, but the defeatee remains defeated too, as will happen whenever a defeater is a universal defeater. Accordingly, any time at which the skeptic believes –R, he has a defeater for –R, even if he also has, at that time, a defeater for that defeater. Skepticism of this kind, then, really is self-defeating, even if it is also the case that the skeptic has a defeater for his defeater.

But the same goes for N&E. At level 0, we have Q (N&E and *P(R/N&E) is low or inscrutable*). Q, however, for any S who believes it, provides S with a (Humean) defeater for R and hence for any other proposition S believes. Let 'Q(p)' denote the belief that p is formed by faculties such that the probability of their reliability on Q is either low or inscrutable. Then perhaps we can represent the relevant structure as follows:

level 0: Q (which is a Humean defeater for R and hence for everything S believes)

level 1: Q and Q(Q), which is a potential defeater for Q
level 2: Q and Q(1), which is a potential defeater for level 1
.
.
level n: Q and Q(n-1), which is a potential defeater for level n-1
.
.
.

Here each level of the structure, if believed, is a defeater for the preceding level; and each level, if believed, is a defeater for itself. The central point, just as with the skeptical structure, is that the reflective naturalist has a defeater for N&E at any time at which he accepts Q. He has such a defeater for N&E even if at that time he also has a defeater for that defeater. His problem, after all, is that N&E and the Probability Thesis give him a defeater for *everything* he believes, including themselves.

Returning to Otte's suggestion, then, the fact that all of the naturalist's beliefs are called into question doesn't mean that he can't rationally conditionalize on any of his beliefs; at the least he must conditionalize on Q. Some of the issues Otte brings up are deep and difficult, and I'm not confident I understand either them or Otte really well. For the moment, however, I prefer to formulate the problem of conditionalization in terms of defeater-deflectors and stick with the formulation I gave above on pages 224–225.

K. To Alston

William Alston puts forward two main objections to EAAN: "(1) Not all the probability assignments are adequately defended," and (and more important) "(2) Even if P(R/N&E) were either low or inscrutable, it would not follow that N&E is a defeater for R, possibly on Plantinga's own conception of defeaters, on one understanding of what that is. At least, I will argue, this doesn't follow on a maximally epistemically significant conception of defeaters. The first step in laying all this out is to consider how Plantinga understands conditional probability. That determination is crucial for both of the above criticisms and also for the evaluation of the inference from the hypothetical population to ourselves."

1. My probability judgments incorrect or unsupported? Alston suggests that the probability of each of the P_i (epiphenomenalism, semantic epiphenomenalism, etc.) on N&E is inscrutable: "With no more than that [N&E] to go on, how could we possibly make even a rough estimate of the likelihood that their beliefs do or do not have a causal influence on behavior . . . ?" I think he is a bit quick here: in particular, P(–C/N&E), the probability of semantic

epiphenomenalism on N&E, seems to me to be high. This is so, at any rate, if we think of materialism with respect to human beings as part of or epistemically very probable with respect to N&E. For (as I argue above) it is extremely hard from that perspective to think of any way in which the content of a belief *could* figure into the etiology of behavior: it is those neurophysiological properties that seem to do the job, and content seems to be irrelevant. Still, EAAN is supposed to work even if P(–C/N&E) is inscrutable; so let's suppose that it is. Inspection of the formula on page 10 reveals that under that condition, if either (or both) P(R/N&E&–C) or P(R/N&E&C) is inscrutable, then P(R/N&E) will also be inscrutable. If neither P(R/N&E&–C) nor P(R/N&E&C) is inscrutable, then P(R/N&E) will fall within the interval bounded by P(R/N&E&–C) and P(R/N&E&C); this will serve in EAAN, since P(R/N&E&–C), if not inscrutable, will be low.

2. *Defeaters.* I argue that P(R/N&E) is low or inscrutable; Alston devotes most of his energy to considering the first possibility. His main criticism "concerns the interpretation and significance of the claim that N&E is a defeater for R"; and perhaps his principal suggestion is that in the situation envisaged in EAAN, we might as well take R (together with the Probability Thesis) to be a defeater for N&E instead of taking N&E (and the Probability Thesis) as a defeater for R: "we need to ask whether this implies that one who holds N&E has a *defeater* for R (rather than vice versa) and if she does whether this makes it irrational for her to hold R (rather than irrational for her to hold N&E)." He begins by examining my account of defeaters and defeat in *Warranted Christian Belief*, which goes as follows:

> (D) *D* is a purely epistemic defeater for *B* for *S* at *t* iff (1) *S*'s noetic structure *N* at t includes *B* and *S* comes to believe *D* at *t*, and (2) any person *S** (a) whose cognitive faculties are functioning properly in the relevant respects, (b) who is such that the bit of the design plan governing the sustaining of *B* in her noetic structure is successfully aimed at truth (i.e., at the maximization of true belief and minimization of false belief) and nothing more, (c) whose noetic structure is *N* and includes *B*, and (d) who comes to believe *D* but nothing else independent of or stronger than *D*, would withhold *B* (or believe it less strongly).[69]

Alston has some good things to say about this definition. In particular, he points out that I say little or nothing about the 'relevant respects' in which S's cognitive faculties are supposed to be functioning properly, and he explores several possibilities as to what I might mean. Right: I should have said what

[69] *Warranted Christian Belief*, 363.

these relevant respects are supposed to be. The main idea is that the *defeater system* is functioning properly; that is the central 'relevant respect'. This means, for example, that the doxastic response to experience is functioning properly ('internal rationality') with respect to D and B and any other beliefs involved in the particular cases of belief formation and sustenance at issue. The module that produces D, on the other hand, need not be functioning properly; it need not be that D is rationally produced or rationally held for it to be a defeater. It is therefore possible for a belief that is acquired or held irrationally to defeat a belief that is held rationally. Second, it also need not be that all the *other* beliefs in S's noetic structure are rational, that is, produced by properly functioning belief-producing processes. This means that I can get a defeater for a rational belief B by rationally acquiring a belief D, which interacts with an irrationally held or acquired belief C in such a way as to be a defeater for B. Thus, for example, I've always thought, rationally, that you tell me only the truth; but I also think, irrationally, that no one lives in Boston; then if I form the rational belief that you have just told me you live in Boston, I may have a defeater for the belief that you always tell me the truth. That an irrationally acquired or held belief can be a defeater of a rationally held belief is important for EAAN because I want to leave open the possibility that N&E is (part of) a defeater for R, even if N&E is not rationally acquired and rationally held; and I want to leave *that* open because it seems to me quite possible that naturalism, with its rejection of belief in God, often or typically arises out of cognitive malfunction—perhaps defective function of the *sensus divinitatis*.

Now I want to argue that the person who accepts N&E and the Probability Thesis has a defeater for R; of course, she could give up N&E, but I say that so long as she accepts N&E and the Probability Thesis she has such a defeater for R. The only way she can achieve alethic rationality, then, is to give up N&E. She can't give up the Probability Thesis without irrationality, if she sees that it is in fact true. As for R, if she gives it up, she acquires a defeater for all that she believes, including, of course, N&E, so that once again rationality requires that she give up N&E. Now I think Alston is inclined to agree with this diagnosis of the situation. However, he has two basic caveats. First, he thinks the naturalist will in fact not see the truth of the Probability Thesis. And second, he thinks that the problem facing the naturalist here can't properly be put in terms of his having a defeater for R.

As for the first, I think a sensible naturalist will in fact be inclined to accept the Probability Thesis. That is because of the great difficulty in seeing how, given naturalism (construed as including materialism), content could possibly be involved in the etiology of behavior, and the obviousness of the fact that if it isn't, the probability of R on N&E will be low or inscrutable. Furthermore, this problem with content—both in seeing how it can accrue to material processes, and in seeing how, if it somehow does, it can enter the causal chain

leading to behavior—this problem with content is part of what underlies the attractiveness of eliminativism for naturalists.[70]

As for the second, Alston's main problem with (D) is just that according to D an irrational belief can defeat a rational belief. Suppose the new belief D is arrived at irrationally; why should we think that it can defeat some belief with impeccable epistemic credentials? There are several different kinds of defeat and defeaters, and there can be sensible debate about what constitutes the most epistemically relevant sense of 'defeat'. Precisely that point is relevant in the present context. Alston thinks an epistemically relevant account of defeaterhood will have the consequence that a defeater D cannot defeat a belief that has more warrant than D does. His picture of the defeating situation, perhaps, goes like this: S acquires a new belief D that conflicts in some way with beliefs he already has. What rationality then requires is that he give up the belief with the least warrant. But then it won't be the case that D defeats a belief that has more warrant than it does.

Now there are several things to be said here. First, a mere technicality: on my picture of belief it is not the case that at some point S (if rational) holds both D, the defeater, and B, the defeatee. Upon acquiring D, she no longer believes B (or believes it less firmly). But this is a mere bagatelle: perhaps the fact is there is a split second where S believes both; or if there isn't, Alston could accommodate this fact by claiming that in a sensible account of defeaters, it will not be the case that S rationally comes to believe a proposition that is in the relevant way incompatible with some other proposition she already believes. More important, however, there is the distinction between internal and external rationality. Internal rationality has to do with proper function, proper cognitive response downstream from experience, including doxastic experience; external rationality has to do with proper function upstream from experience. Now one natural conception of defeaterhood would connect it with internal rationality. Under this conception, a belief that is acquired and held in a way that is externally irrational can nevertheless be a defeater of a belief that is acquired and held rationally, both internally and externally. By virtue of paranoia I come to think you are out to destroy my academic reputation; this gives me a defeater for my previous belief that you are well disposed toward me, even though it is irrationally acquired. Under this conception, defeaterhood has to do with what the design plan requires, *once you acquire a new belief;* it does not have to do with the rationality, for you, of this new belief. The question, given this conception of defeat, is what rationality requires by way of adjustment of belief given that you accept (and continue to hold) this new belief B.

Another example: suppose I am one of Descartes's madmen and come to

[70] And the fact is, eliminativism is one way of acceding to the conclusion of EAAN: for an eliminativist, presumably, won't think of himself as a naturalist—i.e., one who believes naturalism; he doesn't think there is any such thing as believing naturalism.

believe, by virtue of external irrationality, that my head is made of glass. Then I have a defeater, so I say, for my previous, rationally formed belief that it is made of flesh and bone and blood. The question here is this: *given* that I believe that my head is made of glass, what changes in my noetic structure are required by proper function? Well, one thing that will occur (if the defeater system is functioning properly), is that I will no longer believe that my head is made of flesh and bone. Again, I've always thought you were favorably disposed toward me; but then I come, unreasonably, to believe that you are out to get me; if the defeater system is functioning properly, I will not also continue to believe that you are favorably disposed toward me. This seems to me a perfectly respectable and epistemically relevant kind of defeater. Furthermore, it is the one that is relevant in the case of EAAN, at least if we want to leave open the possibility that the belief in naturalism is often irrationally acquired.

Finally, Alston concludes by claiming that "All this leaves Plantinga's argument, as he has presented it, without the intended force. It would still be possible for him to argue that one who insists on holding on to N&E cannot preserve internal rationality except by giving up R. And perhaps that would give him the heart of what he is after. But it doesn't give it to him in the way in which he tried to reach that result." That is indeed the heart of what I am after; it is encouraging to see that Alston agrees on this point, even if he thinks I didn't argue for it properly.

Contributors

William Alston is Professor Emeritus of Philosophy at Syracuse University. His major fields of research are philosophy of religion, epistemology, philosophy of language, and philosophy of mind. His recent books are *Epistemic Justification* (1989), *Divine Nature and Human Language* (1989), *Perceiving God* (1991), *The Reliability of Sense Perception* (1993), *A Realist Conception of Truth* (1996), and *Illocutionary Acts and Sentence Meaning* (2000), all published by Cornell University Press. Most recently his 2001 Aquinas Lecture at Marquette University, *A Sensible Metaphysical Realism*, was published by Marquette University Press.

James Beilby is Assistant Professor of Theology at Bethel College in St. Paul, Minnesota. His primary interests are philosophical theology and religious epistemology. He has published articles in *Faith and Philosophy*, *International Journal for Philosophy of Religion*, *Philosophia Christi*, *Religious Studies*, and other journals. He is also coeditor of *Divine Foreknowledge* (2001), published by InterVarsity Press.

Michael Bergmann is Assistant Professor of Philosophy at Purdue University in West Lafayette, Indiana. His main areas of research are epistemology, metaphysics, and philosophy of religion. He has published in *Philosophical Review*, *Noûs*, *Philosophy and Phenomenological Research*, and other journals.

Evan Fales is Associate Professor of Philosophy at the University of Iowa. Currently his major areas of interest are metaphysics, epistemology, and philosophy of religion. Among his most recent publications are several papers in *Religious Studies* on the evidential value of mystical experience, and critical

studies of Plantinga's *Warranted Christian Belief* (in *Noûs*) and Geivett and Habermas's *In Defense of Miracles* (in *Philosophia Christi*).

Jerry Fodor's major academic appointments have been at MIT, the CUNY Graduate Center, and Rutgers University, where he is currently the State of New Jersey Professor of Philosophy. His main research interests are in cognitive science and the philosophy of mind. Recent books include *Concepts* (Oxford University Press), *In Critical Condition*, and *The Mind Doesn't Work That Way* (both MIT Press).

Trenton Merricks is Professor of Philosophy at the University of Virginia. He is the author of *Objects and Persons* (Oxford University Press, 2001) and many articles in metaphysics and epistemology.

Timothy O'Connor is Associate Professor of Philosophy at Indiana University. He is the editor of *Agents, Causes, and Events* (1995) and the author of *Persons and Causes: The Metaphysics of Free Will* (2000), both with Oxford University Press. He also has written numerous professional articles on topics in metaphysics, philosophy of mind, and philosophy of religion. His two main current interests are the ontology of mind and the metaphysics and epistemology of modality. The latter (with application to the cosmological argument from contingency for theism) is the topic of a book he is now writing.

Richard Otte is Associate Professor of Philosophy at the University of California, Santa Cruz. Although most of his earlier work was in philosophy of science and philosophical logic, recently he has begun working on philosophy of religion. Some of his recent publications are "Evidential Arguments from Evil," in *International Journal for Philosophy of Religion*, and "Rowe's Probabilistic Argument from Evil," in *Faith and Philosophy*.

Alvin Plantinga is the John A. O'Brien Professor of Philosophy at the University of Notre Dame. He is also director of the Center for Philosophy of Religion at Notre Dame and has taught philosophy there for nineteen years; prior to that he taught at Calvin College for nineteen years. His primary areas of research are philosophy of religion, metaphysics, and epistemology. Among his books are *God and Other Minds* (1967), *The Nature of Necessity* (1974), *God, Freedom and Evil* (1974), *Does God Have a Nature?* (1980), *Warrant: The Current Debate* (1993), *Warrant and Proper Function* (1993) and *Warranted Christian Belief* (2000). He spends too much time flying around the United States and other countries, but too little time rock climbing.

William Ramsey is Associate Professor of Philosophy at the University of Notre Dame. His scholarly interests are primarily in the philosophy of cognitive science, particularly the areas of mental representation, philosophical di-

mensions of connectionist modeling, and the psychology of philosophy. He is the coeditor of *Philosophy and Connectionist Theory* (1991) and *Rethinking Intuition* (1998). The former explores the philosophical implications of connectionist cognitive modeling, while the latter examines the psychological and epistemological dimensions of intuition-based philosophy. His most recent project is a book that critically assesses the explanatory value of notions of mental representation in different cognitive disciplines. This work is based in part on his recent article "Do Connectionist Representations Earn Their Explanatory Keep?" which appeared in *Mind and Language* 12, 1 (1997).

Ernest Sosa is Romeo Elton Professor of Natural Theology at Brown University and Distinguished Visiting Professor at Rutgers University. He has published papers in metaphysics and epistemology, and some of his work in epistemology has been collected in *Knowledge in Perspective* (Cambridge University Press, 1991).

W. J. Talbott is Associate Professor of Philosophy at the University of Washington. He has published articles in epistemology, moral and political philosophy, rational choice theory, philosophy of law, and the theory of self-deception. He is currently working on a book in epistemology, to be titled *Learning from Experience*, and a book in political philosophy, to be titled *Why Human Rights Should Be Universal.*

James Van Cleve is Professor of Philosophy at Brown University. He is the author of *Problems from Kant* (Oxford University Press, 1999) as well as many articles on topics in epistemology and metaphysics.

Index

a priori knowledge, 67, 82–86, 88
abstract faculties, 163, 164
adaptive behavior, 4, 16, 19–20, 41
Alston, William, 88n, 105n, 110, 271
animal knowledge, 96n, 99, 108–110, 238, 247
Aristotle, 96

Bayes' theorem, 120n, 137–139, 142
behavior-producing mechanisms, 219
Beilby, Jim, 1, 204
beliefs,
  adaptive, 8
  basic, 66, 78, 81
  causally equivalent, 7n
  commonsense, 67, 82
  logically equivalent, 7n
  maladaptive, 8–9
  memory, 69, 238
  noninferential, 50, 66–67, 87–88, 230
  semantic properties of, 7, 9, 16, 212, 257, 262–263
  sense perceptual, 67, 81, 88, 145, 160, 162, 238
  syntactical properties of, 7, 212, 262–263
  theoretical, 162
  truth value of a, 16, 18, 47, 49
belief-desire pairs, 22, 136, 147, 260
belief-forming mechanisms, 6, 16, 19, 22, 24, 27, 186, 219
  reliable, 46, 154
  unreliable, 22–23
Bergmann, Michael, 230, 232

Calvin, John, 266
Cartesian circle, 93, 103, 238
causal theory of reference, 27
Chisholm, R., 85
Churchland, Patricia, 3–4, 32
clear and distinct perception, 105–107, 110–113, 241
closure principle, 112, 114
cognitio, 108, 111, 238
cognitive faculties, 2–4, 46, 143, 149, 184, 207
  reliable, 137
cognitive mechanisms, 21–22, 25, 32, 43
  reliable, 44
  unreliable, 20
common sense knowledge, 83–89
conative mechanism, 32
conditional probability, 135, 142, 167, 175, 178, 271
conditionalization problem, viii, 220–221, 240
confirmation theory, 138, 141
"contextualist" response, 112n
Cruz, Joseph, 85,
Cummins, Robert, 7, 10

Darwin, Charles, 3–5, 29
Darwin's doubt, 10, 46
Darwinism, 39, 41
  neo-, 45, 47
Davidson, Donald, 5
defeaters, vii, ix, 2, 11, 62, 69, 73, 108, 112, 116–117, 141, 155, 174–175, 194, 196–197, 200, 207, 243

defeaters *(continued)*
candidate, 194
epistemic, 187
Humean rationality, 211, 220, 225, 228, 231, 235–236, 239–240, 243, 245
prima facie, 194
probabilistic, 224
proper-function rationality, 209, 211, 225, 228, 231, 236, 244–245
purely alethic rationality, 209, 211, 225, 228, 231, 236, 244–245
rationality, 187, 205, 208–209
rebutting, 70, 72, 193
undercutting, 70, 72, 153–154, 162, 185n, 193
universal, 269–270
warrant, 187
Dennett, Daniel, 41
DeRose, Keith, 105n, 107–108, 110, 113
Descartes, René, 57, 92, 95, 98, 113, 116
description theory of reference, 27
design, 3n
intelligent, 28
desire-forming mechanism, 22
Devitt, Michael, 61n, 84
direct inference, 159
Donnellan, Keith, 27
Dretske, Fred, 7n, 165n, 213–214
dualism, 212n
Cartesian, 214
substance, 220

eliminativism, 274n
emergent properties, 216
epiphenomenalism, 6, 9, 182, 262, 271
semantic, 6, 9–10, 211–212, 214–215, 253, 257–258, 262, 271
epistemic circularity, 79, 82, 110, 200, 226
epistemically circular argument, 77–78, 241–242
evolution, 2, 12, 98, 116, 156
contemporary theory of, 1, 5, 15, 149, 186
evolutionary epistemology, 28
evolutionary reliabilism, 16, 19–20, 27, 29
externalism, 47, 79n, 96, 104–105, 111, 124
externalist epistemology, 96

Fales, Evan, 3n, 262
Falk, Arthur, 51n
false beliefs, 19, 214–215, 260
adaptive, 41
Feldman, Richard, 23
Fodor, Jerry, 17n, 216
foundationalism, 63, 93, 96
Cartesian, 53
Freud, Sigmund, 185, 209

Garcia effect, 24–25, 29
Garcia, J., 24n
genetic drift, 2
Gettier-proof theory, 43
Ginet, Carl, 221–222
God, 111, 137, 139
Godel's incompleteness theorem, 163n
Goldman, Alvin, 50n, 58n, 83n
Gould, S. J., 40n

Hare, John, 5
Heller, Mark, 105n, 111, 113
Hempel, Carl, 41n
Hume, David, 210, 269
Huxley, T. H., 6n

implicit premise thesis, 122, 240–241, 244
inductive inferences, 57
inference strategies, 22, 24, 28
inferential mechanisms, 49–50
internalism, 79n, 96

Kitcher Philip, 84
Kripke, Saul, 26

Lewis, C. S., ix
Lewis, David, 161
Lewontin, R. C., 40n
logically circular arguments, 77–78

materialism, 10, 213–214, 272, 273
Milton, Katherine, 48n
Mirza, Omar, 3n
Moorean response, 112–114

natural selection, 2, 4, 7, 19–20, 44, 136, 185–186, 214, 257
naturalism, 1–4, 12, 28, 33, 62, 87, 98, 116, 138, 149, 223, 243, 252, 261, 265, 273
epistemological, 83–84
metaphysical, vii, 61, 83–85, 87, 212
moderate, 83
ontological, 43
philosophical, 1, 3
scientistic, 83
naturalized epistemologies, 43–45
theistic, 44, 52
nonpropositional evidence, 87
Nozick, Robert, 165
"Nozickian" response, 112

O'Connor, Timothy, 223, 234
Otte, Richard, 220

particularism, 85
particularist methodology, 88
Perry, John, 221–223
physicalism, 84

Plantinga, Alvin, 45, 69n, 88n, 116, 122, 129, 165
  epistemology of, 191–192
  naturalized epistemology of, 43
Pollock, John, 11n, 85, 153
Popper, Karl, 185
Prichard, H. A., 114
principle of epistemic ascent, 109
principle of exclusion, 109
principle of transfer of justification, 109
probabilistic argument from evil, 63–64
probabilistic inference, 159
probability thesis, 205, 223, 235, 240, 243, 272
proper function, 129, 131, 188, 205–206, 235, 251
  cognitive, 228, 232
Putnam, Hilary, 30, 50n

Quine, W. V. O., 185, 222

Ramsey, William, 256
random genetic mutation, 2
random genetic variation, 3n
Rea, Michael, 261
reflective knowledge, 91, 94–96n-97, 99, 101, 108–110, 113, 238, 246–247
Reid, Thomas, 2, 61, 64n, 87, 101, 230, 242
Reidian position, 124
reliabilism, 95, 98n
Ruse, Michael, 260
Russell, Bertrand, 251, 260

scientia, 108, 111n, 115, 238, 246
scientific realism, 138
scientism, 30, 33, 39, 251–252
Sextus Empiricus, 269
Simon, Herbert, 57
skepticism, viii, 39, 44, 89, 125, 238
  ancient, 95, 243
  global, 62, 96
  modern, 95
Smith, J. M., 6
Sober, Eliot, 18n
Sosa, Ernest, 105, 107, 113, 238, 243
Stich, Stephen, 28n, 45

Talbott, William, 206, 256
Taylor, Richard, ix
theism, vii, 44, 52, 54, 123, 133, 137–139, 221, 223, 237, 266
  austere, 119, 221
theistic reliabilism, 29
theoretical inferences, 58
Thomas Aquinas, 2n, 266
true beliefs, 3, 109, 208, 215, 232, 235, 238, 251, 257
  adaptive, 41

Van Cleve, James, 238

warrant, 108, 121, 131, 173, 175, 190–192, 207, 235, 250
Wilson, E. O., 260
Wisniewski, Edward, 86–87n

PHILOSOPHY/METAPHYSICS

Almost a decade ago, Alvin Plantinga articulated his bold and controversial evolutionary argument against naturalism. This intriguing line of argument raises issues of importance to epistemologists and to philosophers of mind, of religion, and of science. In this, the first book to address the ongoing debate, Plantinga presents his influential thesis and responds to critiques by distinguished philosophers from a variety of subfields.

Plantinga's argument is aimed at metaphysical naturalism or roughly the view that no supernatural beings exist. Naturalism is typically conjoined with evolution as an explanation of the existence and diversity of life. Plantinga's claim is that one who holds to the truth of both naturalism and evolution is irrational in doing so. More specifically, because the probability that unguided evolution would have produced reliable cognitive faculties is either low or inscrutable, one who holds both naturalism and evolution acquires a "defeater" for every belief he/she holds, including the beliefs associated with naturalism and evolution.

Following Plantinga's brief summary of his thesis are eleven original pieces by his critics. The book concludes with a new essay by Plantinga in which he defends and extends his view that metaphysical naturalism is self-defeating.

JAMES BEILBY is Assistant Professor at Bethel College in St. Paul, Minnesota.

Cornell Paperbacks
Cornell University Press
www.cornellpress.cornell.edu

Univ. of Notre Dame
NATURALISM DEFEATED?
000853328
000853328
Auth:
SubD:PHILOSOPHY
4/04/03
$19.95
Class:PHILOSOPHY-GENE
SubC:PHILOSOPHY